

Suresh Chandra Satapathy
P. S. Avadhani
Ajith Abraham (Eds.)

**Proceedings of the International
Conference on Information
Systems Design and Intelligent
Applications 2012 (INDIA 2012)
held in Visakhapatnam, India,
January 2012**

Advances in Intelligent and Soft Computing

Editor-in-Chief

Prof. Janusz Kacprzyk
Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6
01-447 Warsaw
Poland
E-mail: kacprzyk@ibspan.waw.pl

Further volumes of this series can be found on our homepage: springer.com

Vol. 116. Yanwen Wu (Ed.)
Advanced Technology in Teaching - Proceedings of the 2009 3rd International Conference on Teaching and Computational Science (WTCS 2009), 2012
ISBN 978-3-642-11275-1

Vol. 117. Yanwen Wu (Ed.)
Advanced Technology in Teaching - Proceedings of the 2009 3rd International Conference on Teaching and Computational Science (WTCS 2009), 2012
ISBN 978-3-642-25436-9

Vol. 118. A. Kapeczynski, E. Tkacz,
and M. Rostanski (Eds.)
Internet - Technical Developments and Applications 2, 2011
ISBN 978-3-642-25354-6

Vol. 119. Tianbiao Zhang (Ed.)
Future Computer, Communication, Control and Automation, 2011
ISBN 978-3-642-25537-3

Vol. 120. Nicolas Loménie, Daniel Racoceanu,
and Alexandre Gouaillard (Eds.)
Advances in Bio-Imaging: From Physics to Signal Understanding Issues, 2011
ISBN 978-3-642-25546-5

Vol. 121. Tomasz Traczyk and
Mariusz Kaleta (Eds.)
Modeling Multi-commodity Trade: Information Exchange Methods, 2011
ISBN 978-3-642-25648-6

Vol. 122. Yinglin Wang and Tianrui Li (Eds.)
Foundations of Intelligent Systems, 2011
ISBN 978-3-642-25663-9

Vol. 123. Yinglin Wang and Tianrui Li (Eds.)
Knowledge Engineering and Management, 2011
ISBN 978-3-642-25660-8

Vol. 124. Yinglin Wang and Tianrui Li (Eds.)
Practical Applications of Intelligent Systems, 2011
ISBN 978-3-642-25657-8

Vol. 125. Tianbiao Zhang (Ed.)
Mechanical Engineering and Technology, 2011
ISBN 978-3-642-27328-5

Vol. 126. Khine Soe Thuang (Ed.)
Advanced Information Technology in Education, 2011
ISBN 978-3-642-25907-4

Vol. 127. Tianbiao Zhang (Ed.)
Instrumentation, Measurement, Circuits and Systems, 2011
ISBN 978-3-642-27333-9

Vol. 128. David Jin and Sally Lin (Eds.)
Advances in Multimedia, Software Engineering and Computing Vol.1, 2011
ISBN 978-3-642-25988-3

Vol. 129. David Jin and Sally Lin (Eds.)
Advances in Multimedia, Software Engineering and Computing Vol.2, 2011
ISBN 978-3-642-25985-2

Vol. 130. Kusum Deep, Atulya Nagar,
Millie Pant, and Jagdish Chand Bansal (Eds.)
Proceedings of the International Conference on Soft Computing for Problem Solving (SOCPROS 2011) December 20–22, 2011, 2012
ISBN 978-81-322-0486-2

Vol. 131. Kusum Deep, Atulya Nagar,
Millie Pant, and Jagdish Chand Bansal (Eds.)
Proceedings of the International Conference on Soft Computing for Problem Solving (SOCPROS 2011) December 20–22, 2011, 2012
ISBN 978-81-322-0490-9

Vol. 132. Suresh Chandra Satapathy, P.S. Avadhani,
and Ajith Abraham (Eds.)
Proceedings of the International Conference on Information Systems Design and Intelligent Applications 2012 (INDIA 2012) held in Visakhapatnam, India, January 2012, 2012
ISBN 978-3-642-27442-8

Suresh Chandra Satapathy, P.S. Avadhani,
and Ajith Abraham (Eds.)

Proceedings of the International
Conference on Information
Systems Design and Intelligent
Applications 2012 (INDIA 2012)
held in Visakhapatnam, India,
January 2012



Springer

Editors

Dr. Suresh Chandra Satapathy
Andhra University
Dept of Computer Science
and Engineering
ANITS, Sangivalasa
530003 Vishakapatnam
India
E-mail: sureshsatapathy@gmail.com

Prof. Ajith Abraham
Machine Intelligence Research Labs
Auburn, WA USA
E-mail: ajith.abraham@ieee.org

Dr. P.S. Avadhani
Andhra University
College of Engineering
Dept. of CS&SE
ANITS, Sangivalasa
530003 Vishakapatnam
India
E-mail: psavadhani@yahoo.com

ISBN 978-3-642-27442-8

e-ISBN 978-3-642-27443-5

DOI 10.1007/978-3-642-27443-5

Advances in Intelligent and Soft Computing

ISSN 1867-5662

Library of Congress Control Number: 2011944253

© 2012 Springer-Verlag Berlin Heidelberg

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Typeset by Scientific Publishing Services Pvt. Ltd., Chennai, India

Printed on acid-free paper

5 4 3 2 1 0

springer.com

Preface

This AISC volume contains the papers presented at the First International Conference on Information Systems Design and Intelligent Applications (INDIA-2012) with a theme 'Role of India Inc in Global Scenario' held during 5–7 January 2012 organized by Computer Society of India, Vishakhapatnam Chapter in association with Vishakhapatnam Steel plant, a Navaratna Company.

INDIA-2012 is the beginning of the prestigious international conference series that is targeted to bring researchers from academia and industry to report, deliberate and review the latest progresses in the cutting-edge research pertaining to intelligent computing, data engineering, networking among few others.

We received research submissions in the various advanced technology areas and after a rigorous peer-review process with the help of our program committee members and external reviewers finally we accepted papers with an acceptance ratio of 0.48.

The conference featured many distinguished keynote address by eminent speakers like Dr. Ajith Abraham, Technical University of Ostrava, Czech Republic, Dr. T.R. Gopalakrishnan Nair, RIIC, Bangalore, Dr. Louise Perkins, University of Southern Mississippi, Dr. Sumanth Yenduri, University of Southern Mississippi, Dr. Sanghamitra Bandyopadhyay, ISI Kolkota. We also had galaxy of speakers from industries such as Mr. M. Moni, DDG, NIC, New Delhi, Sri U. Rama Mohan Rao, SP, CID, Cybercrimes, Hyderabad, Mr. Venkatesh Kallu, NouveauGEN Solutions, Canada.

We take this opportunity to thank authors of all submitted papers for their hard work, adherence to the deadlines and patience with the review process. The quality of a referred volume depends mainly on the expertise and dedication of the reviewers. We are indebted to the program committee members and external reviewers who not only produced excellent reviews but also did these in short time frames.

We would also like to thank CSI, Visakhapatnam Chapter having coming forward to organize this first ever conference in the series. Computer society of India (CSI) was established in 1965 with a view to increase information and technological awareness among Indian society, and to make forum to exchange and share the IT-related issues. The headquarters of the CSI is situated in Mumbai with a full-fledged office setup and is coordinating the individual chapter activities. It has 70 chapters and more than 400 students' branches operating in different cities of India. The total strength of CSI is above 70000 members. CSI, Visakhapatnam chapter is having more than two decade long history. This chapter is one of most active in CSI-INDIA. All the top executives of the industry and academia are life members in this chapter. CSI-Vizag developed and maintains its website www.csi-vizag.org which is regularly updated with a treasure of IT information and presentations. It has felicitated the doyens of industries and at the same time always encouraged the students in various forms. It has conducted many IT-awareness programmes at Regional, National and

International level Conferences/Seminars for members. Vishakhapatnam Chapter maintains eNews quarterly and also provides resource persons to Student Branches on Quarterly basis free of charge from Academia and Industries. To give a fillip to the Computer Science and IT students, CSI-Vizag has constituted a first of its kind CSI award since 2010 for the meritorious students 'CSI Meritorious students awards' and gave away the awards to the toppers of the Universities within the chapter (Srikakulam, Visakhapatnam, VZM, EGDist, WGDist these five districts covered this chapter). CSI-Vizag also started 'Best Performance Awards' to the Student Chapters to encourage their active performance since 2010. For all these activities this chapter was adjudged with BEST Chapter award for two times in the years 2005 & 2010.

We are indebted to Visakhapatnam Steel plant (Always Back bone to CSI-Vizag in all activities) for their immense support to make this international conference possible in such a grand scale. Visakhapatnam Steel Plant (VSP), a Govt. of India Undertaking under the corporate entity of Rashtriya Ispat Nigam Ltd., is the first shore-based integrated steel plant in India. The plant with a capacity of 3 mtpa was established in the early nineties and is a market leader in long steel products (wire rods, rebars, rounds, angles, channels, blooms and billets) catering to the construction, automobile, wire drawing, forging and other manufacturing segments. The Plant is almost doubling its capacity to a level of 6.3 mtpa of liquid steel at a cost of around 2500 million USD and the products from the new units are set to come on stream from end of 2011-12 progressively. The plant has been operating consistently beyond its rated capacities in the range of 120% for more than a decade with a turnover of over 2400 million USD consecutively for the last four years. The plant has an excellent layout, infrastructure, logistics and adequate land bank for expansion to 20 mtpa. The presence of a deep draft port in the vicinity offers unique advantage of unloading Cape vessels and conveyorized transfer to the Plant. RINL-VSP is the first integrated steel plant in India to be accredited with all three international standards, viz. ISO 9001, ISO 14001 and OHSAS 18001. It is also the first Steel Plant to be certified with CMMI level-3 certificate and BS EN 16001 standard. The plant though basically designed for production of mild steel grades, today produces nearly 80% of its saleable steel as value added steel, through in-house innovative process improvements. In a study conducted by CRU, London, RINL-VSP was adjudged as 5th lowest cost long steel producer in the world.

Our special thanks to all past chairmen & his MC & NC members of CSI-Vizag & CSI-India, Sri Ravindra Ranjan, Sri P.C. Mohapatra, Sri P. Ramudu, Sri G.V.N. Reddy, Sri M.M.K. Murthy, Sri P. Moharikar, Sri G.N. Murthy, Sri G. Jogeswara Rao, Sri Y Sudhakar Rao, Sri T.V.S. Krishna Kumar, Sri K. Iyapilla, Sri P. BalaChandra Rao, Sri R. Bhaskar of Vizag Steel for their continuous support to CSI-Vizag. We are also thankful to Sri L. Bhaskar, Sri Deepankar Das, Sri Anzar Alam, Sri DVS Kumar and staff members of CMD, D(O), D(F) & D(Proj) offices of Vizag Steel for their continuous support to CSI in all activities. Thanks to all CSI Student branch coordinators, Administration & Management of Engg. colleges under Visakhapatnam chapter for their continuous support to our chapter activities. Sincere thanks to employees of Vizag steel, CSI-Vizag members those who are supported CSI-Vizag activities directly or indirectly.

Our thanks are due to Sri V. Thapovardhan, Secretary and Correspondent of ANITS and Prof. V.S.R.K. Prasad, Principal ANITS for their encouragement and for making the services of Dr. Suresh Chandra Satapathy and his team available for enriching the academic activities of this conference.

Our sincere thanks to local industries Management and sponsors to support us in all activities and made this conference a grand success.

Our sincere thanks to all the chairs who have guided and supported us from the beginning of the inception of the idea of such conference. We extend our heart-felt thanks to Prof. Siba K. Udgata, University of Hyderabad, for his immense cooperation in preparing the entire proceeding related matters. Special thanks to local organizing committee members from Vizag steel & ANITS Engineering College. We would also like to thank the participants of this conference, who have considered the conference above all hardships. Finally, we would like to thank all the volunteers from Vizag steel, ANITS, MVGR, GITAMIT, Avanthi Group of Institutions, Raghu, LENDI, GIET, Chaitanya, Sri Vasavi, BVC, Pragathi, AITAM, GMRIT, DIET Engineering Colleges who spent tireless efforts in meeting the deadlines and arranging every detail to make sure that the conference can run smoothly. All the efforts are worth and would please us all, if the readers of this proceedings and participants of this conference found the papers and conference inspiring and enjoyable. Our sincere thanks to senior life members, life members, associate life members and student members of our Chapter for their cooperation and support for all activities.

Our sincere thanks to all press print & electronic media for their excellent coverage of this conference.

Volume Editors

January 2012

Suresh Chandra Satapathy
P.S. Avadhani
Ajith Abraham

Organization

Chief Patron

Sri A.P. Choudhary

Chairman, CSI-Vizag & CMD, RINL, Visakhapatnam
Steel Plant, Visakhapatnam, AP, India

Patrons

Sri M.D. Agrawal
Sri UmeshChandra

President, CSI-India
Immed.Past Chairman, CSI-Vizag Director (Operations),
RINL, Visakhapatnam, AP, India

Sri P. Madhusudan

Vice-Chairman, CSI-Vizag & Director (Finance), RINL,
Visakhapatnam, AP, India

Honorary Chairs

Dr. Jacek M. Zurada
Dr. V.Bhujanga Rao
Sri C.S. Rao

University of Louisville, Louisville, Kentucky, USA
Chief Controller(R&D), DRDO, New Delhi
President, Corporate Affairs & Regulatory,
Reliance Comm. Ltd, New Delhi

Sri Satish Babu
Prof. P.V.G.D. Prasad
Reddy

Vice President, CSI-India
Rector, AU, Waltair, AP, India

Prof. Allam AppaRao
Prof. Thrimurthy

Ex VC, JNTUK, KKD, AP, India
Immediate Past Chairman, CSI, India

General Chairs

Dr. Narayana C. Debnath
Prof. H.R. Vishwakarma
Dr. G.S.N. Raju
Dr. D. RadhaKrishna
Dr. D.B.V. Sarma
Dr. K. RajaSekhar Rao

PhD, DSc, Winona State University, USA
Secretary, CSI-India, India
Principal, AUCE, India
Principal, AUCE for Women, India
Vice-President, Region-V, CSI, India
Principal, KL University, India

Advisory Board

Sri Ajay Kallam	IAS, Chairman, Visakhapatnam Port Trust
Sri T.K. Chand	Director (Commercial), RINL, VSP Steel Plant
Sri Y.R. Reddy	Director(Personnel), RINL, VSP Steel Plant
Rear Admiral N.K. Mishra	NM,IN (Retd.), CMD, HSL, Vizag
Sri S.V. Ranga Rajan	Director, NSTL
Sri NagaRaj	ED, BHPV
Sri P.A.B. Raju	ED , HPCL
Sri R.C. Swain	Vice President & Head Projects, Jindal aluminium
Sri D.K. Sood	General Manager I/C, NTPC Simhadri
Sri Ch. Prabhakara Rao	General Manager, HZL
Prof. V. Ravindra	Registrar, JNT University, Kakinada, E.G. Dist.
Prof. G. Subramanyam	Vice Chancellor, GITAM University, VSP
Prof. S.R. Gollapudi	GITAM Univesity, Convenor, Advisory Board INDIA-2012

Publicity Chairs

Dr. Somanath Tripathy	IIT, Patna, India
Dr. Naem Hannon	Multimedia University, Malaysia
Dr. A.Janaki Ram	IIT, Madras, India

Program/Technical Committee

Dr. Lydia Ray, USA
Dr. Chaoyang Zhang, Director, School of Computing, USM, USA
Dr. Hamid Arabnia, University of Georgia, USA
Dr. Wei Ding, USA
Dr. J.V.R. Murthy, JNTUK, India
Dr. A. Damodaram, JNTUK, India,
Dr. M.Sashi, AU, India
Dr. Pallam Setty, AU, India
Dr. V. ValliKumari, AU, India
Dr. V. KamakshiPrasad, JNTUH, India
Dr. V. Sree Hari Rao, JNTUH, India
Dr. B.B. Mishra, SIT, India
Dr. V.M. Shenbagraman, SRM University, India
Dr. R. Selva Rani, India
Dr. Amit Saxena, GGUCU, India
Dr. S.S. Gantayant, GIET, Inida
Dr. V. Nagalakshmi, GITAM, India
Dr. S. Udgata, Hyderabad Central University, India
Dr. S. Srinivasa Rao, MVGR, India

Program Chairs

Dr. P.S. Avadhani, Andhra University, India

Dr. S.C. Satapathy, ANITS, India

Sri P. Chandra Sekhar, Vizag Steel, India

Organizing Committee

Vice-Chairman	Sri K.V.S.S. Rajeswara Rao	DGM(IT),VSP
Organizing Secretary	Sri Paramata Satyanarayana	Sr.Mgr(IT),VSP
Finance Secretary	Sri J.V. Rao	AGM(Const),VSP
Website In-charge	Sri Sudhansu Choudhary	AGM(IT), VSP

Local Organizing Committee

Sri Dr. S.N. Rao	Sri Debashish Ray	Sri Y. Arjun Kumar
Dr. V. Phaneendrudu	Sri Suman Das	Dr. B.G. Reddy
Sri G. Thyaga Raju	Sri P. Srinivasulu	Sri M.S. Babu
Sri G.V. Ramesh	Sri S.K. Mishra	Sri V.D. Awashtee
Sri S. Gopal	Sri P.M. Divecha	Sri Y.N. Reddy
Sri B. Satyendra	Sri C.K. Padhi	Sri Uday Kumar
Sri Peddada Srinivas	Sri D.V.G.A.R.G. Varma	Sri P. Sesha Srinivas
Sri A. Paul	Sri V. Hema Sundara Rao	Sri T.N. Sanyasi Rao
Sri P. Balaramu	Sri Shailendra Kumar	Sri Y. Madhu Sudan Rao
Sri G.V. Saradhi	Smt T. Kalavathi	Sri A.P. Sahu
Sri P.M.M. Jayananda	Sri S. Adhinarayana	Sri V. Ratnakar
Sri B.V. Vijay Kumar	Sri Bh. B.V.K. Raju	Sri K. Ravi Ram
Sri M.M.K. Gandhi	Sri Sri Rama Murthy	Sri K. Satyanarayana
Sri S. Raja of NSTL	Sri B.V. Rao of MicroCare	Dr. K.V.L. Raju
Dr. S. Srinivasa Rao	Prof. V. Naga Lakshmi	Dr. C. Mohan Rao
Sri T. Srinivasa Rao	Mr. B. Tirumula Rao	Ch. Suresh
Mr. James Stephen	Mr. Y.V.S.M. Murthy	

Contents

Optimization of Task Allocation Using Quantum Game Theory with Artificial Intelligence	1
<i>G.R. Brindha, S. Anand, S. Prakash, P.M. JoePrathap</i>	
Improving the Grid Scheduling Performance with Fault Tolerance Using Genetic Algorithm	11
<i>Minu Jacob, Sathya Lakshmi, Roberts Masilamani</i>	
The Performance Analysis of a Novel Enhanced Artificial Bee Colony Inspired Global Best Harmony Search Algorithm for Clustering	21
<i>V. Krishnaveni, G. Arumugam</i>	
Artificial Bee Colony Based Image Clustering	29
<i>Kalyani Manda, Suresh Chandra Satapathy, K. Rajasekhara Rao</i>	
Novel Approach to Polygamous Selection in Genetic Algorithms	39
<i>Rakesh Kumar, Jyotishree</i>	
Gang Scheduling Strategy for Request Processing in Cluster Based Video-on-Demand Systems	47
<i>A. Vinay, K. Bharath, T.N. Anitha</i>	
Does Social Communicability Mediate the Role of Trust in Mobile Phone Adoption? An Individual Level Multi-nation Exploratory Study	59
<i>Kallol Bagchi, Somnath Mukhopadhyay</i>	
Data Security for Virtual Data Centers by Commutative Key	67
<i>Maram Balajee, Challa Narasimham, Y. Ramesh Kumar</i>	
Genetic Algorithm for Optimizing Functional Link Artificial Neural Network Based Software Cost Estimation	75
<i>Tirimula Rao Benala, Satchidananda Dehuri, Suresh Chandra Satapathy, S. Madhurakshara</i>	

Detecting and Searching System for Event on Internet Blog Data Using Cluster Mining Algorithm	83
<i>Robin Singh Bhadoria, Manish Dixit, Rohit Bansal, Abhishek Singh Chauhan</i>	
Knowledge Based Evolutionary Programming: Cultural Algorithm Approach for Constrained Optimization	93
<i>Bidishna Bhattacharya, Kamal Mandal, Niladri Chakraborty</i>	
New Measure of Interestingness for Efficient Extraction of Association Rules	103
<i>Parvati Bhurani, Mushtaq Ahmed, Yogesh Kumar Meena</i>	
Improving Prediction of Interdomain Linkers in Protein Sequences Using a Consensus Approach	111
<i>Piyali Chatterjee, Subhadip Basu, Mahantapas Kundu, Mita Nasipuri</i>	
Unconstrained Handwritten Digit OCR Using Projection Profile and Neural Network Approach	119
<i>Amit Choudhary, Rahul Rishi, Savita Ahlawat</i>	
Reduct Generation by Formation of Directed Minimal Spanning Tree Using Rough Set Theory	127
<i>Asit Kumar Das, Shampa Sengupta, Saikat Chakrabarty</i>	
Misclassification and Cluster Validation Techniques for Feature Selection of Diseased Rice Plant Images	137
<i>Santanu Phadikar, Asit Kumar Das, Jaya Sil</i>	
A Novel GA-SVM Based Multistage Approach for Recognition of Handwritten Bangla Compound Characters	145
<i>Nibaran Das, Kallol Acharya, Ram Sarkar, Subhadip Basu, Mahantapas Kundu, Mita Nasipuri</i>	
Cascaded Correlation Neural Network Based Microcalcification Detection in Mammographic Images	153
<i>J. Dheebea, S. Tamil Selvi</i>	
TertProt: A Protein Fold Recognition Method Using Protein Secondary Structure Program	161
<i>D.S.V.G.K. Kaladhar</i>	
Analysis of <i>E.coli</i> Promoter Regions Using Classification, Association and Clustering Algorithms	169
<i>D.S.V.G.K. Kaladhar, T. Uma Devi, P.V. Lakshmi, R. Harikrishna Reddy, R.K. SriTeja Ayayangar V, P.V. Nageswara Rao</i>	
A Parameterization Study of Short Read Assembly Using the Velvet Assembler	179
<i>Alex Christopher Elliot, A. Louise Perkins, Sumanth Yenduri</i>	

A Novel Digital Video Watermarking Scheme Based on the Scene Change Analysis	187
<i>Tummalapalli Geetamma, T.V.N.N.M. Vamsi Krishna, D. Srinivasa Rao</i>	
Application of Swarm Intelligence Computation Techniques in PID Controller Tuning: A Review	195
<i>Soumya Ghosal, Rajkumar Darbar, Biswarup Neogi, Achintya Das, Dewaki N. Tibarewala</i>	
A Steganographic Scheme for Color Image Authentication Using Z-Transform (SSCIAZ)	209
<i>Nabin Ghoshal, Soumit Chowdhury, Jyotsna Kumar Mandal</i>	
Heart Disease Diagnosis Using Machine Learning Algorithm	217
<i>Shashikant U. Ghumbre, Ashok A. Ghatol</i>	
Position Determination and Face Detection Using Image Processing Techniques and SVM Classifier	227
<i>Gogula Suvarna Kumar, P.V.G.D. Prasad Reddy, Sumit Gupta, Ravva Anil Kumar</i>	
An Experimental Analysis of Phylogenetic Trees Based on Topological Score	237
<i>Manoj Kumar Gupta, Rajdeep Niyogi, Manoj Misra</i>	
A Novel Full Reference Image Quality Index for Color Images	245
<i>Prateek Gupta, Priyanka Srivastava, Satyam Bhardwaj, Vikrant Bhateja</i>	
An Energy Efficient On-Demand Routing by Avoiding Voids in Wireless Sensor Network	255
<i>J.D. Preethi, R. Sumathi</i>	
GP Boosting Classification on Concept Drifting Data Streams	265
<i>Dirisala J. Nagendra Kumar, J.V.R. Murthy, Suresh Chandra Satapathy, S.V.V.S.R. Kumar Pallela</i>	
Intelligent Chaos Controller: A Computational Intelligence Based Approach for Data-Driven Real-World Systems	273
<i>Jallu Krishnaiah, C.S. Kumar, M.A. Faruqi</i>	
Protein Structure Prediction in 2D HP Lattice Model Using Differential Evolutionary Algorithm	281
<i>Nanda Dulal Jana, Jaya Sil</i>	
Neighborhood Search Operator Tuned Differential Evolution for Solving Non Convex Economic Dispatch Problem	291
<i>J. Jasper, T. Aruldoss Albert Victoire</i>	

Performance Evaluation of Noise Subspace Methods of Frequency Estimation Techniques	299
<i>Kakaraparathi Bramaramba, S. Koteswara Rao, K. Raja Rajeswari</i>	
Performance Analysis of DSSS System Using Adaptive Filters in Interference Prone Environment	309
<i>Katyayani Kaligathi, Kandula Srinivasa Rao, Seetala Santha Kumari, G. Prabhakara Rao</i>	
An Improved Order Estimation of MSF for Stereophonic Acoustic Echo Cancellation	319
<i>Asutosh Kar, Alaka Barik, Ravinder Nath</i>	
Application of Stochastic Model on Routing Technique in Multi Class Queueing Network	329
<i>K. Sivaselvan, C. Vijayalakshmi Seshathri</i>	
An Efficient Data Structure for Document Clustering Using K-Means Algorithm	337
<i>Ramanji Killani, Suresh Chandra Satapathy, A.M. Sowjanya</i>	
Unconstrained Optimization for Maximizing Ultimate Tensile Strength of Pulsed Current Micro Plasma Arc Welded Inconel 625 Sheets	345
<i>Kondapalli Siva Prasad, Y.V. Srinivasa Murthy, Ch. Srinivasa Rao, D. Nageswara Rao, Gurrala Jagadish</i>	
Simple and Effective Techniques for Skew Correction, Slant Correction and Core-Region Detection for Cursive Word Recognition	353
<i>Kota Virajitha, B. Navya, L.N. Phaneendra Kumar Boggavarapu, Radhe Syam Vaddi, Hima Deepthi Vankayalapati</i>	
A New CDMA Based 2.5G Network Base Station Power Control Algorithm for Improving User Capacity	363
<i>Ramarakula Madhu, G. Sasi Bhushana Rao</i>	
Performance Evaluation of the Gateway Discovery Approaches under Varying Node Speed	371
<i>Koushik Majumder, Sudhabindu Ray, Subir Kumar Sarkar</i>	
Speech Enhancement and Recognition of Compressed Speech Signal in Noisy Reverberant Conditions	379
<i>Maloji Suman, Habibulla Khan, M. Madhavi Latha, Devarakonda Aruna Kumari</i>	
Emission Constrained Economic Dispatch Using Logistic Map Adaptive Differential Evolution	387
<i>Kamal K. Mandal, Bidishna Bhattacharya, Bhimsen Tudu, N. Chakravorty</i>	

PSO Based Edge Keeping Suppression of Impulses in Digital Imagery	395
<i>Jyotsna Kumar Mandal, Somnath Mukhopadhyay</i>	
Non-recursive FIR Band Pass Filter Optimization by Improved Particle Swarm Optimization	405
<i>Sangeeta Mandal, Sakthi Prasad Ghoshal, Rajib Kar, Durbadal Mandal, S. Chaitnya Shiva</i>	
Text Categorization with K-Nearest Neighbor Approach	413
<i>Suneetha Manne, Sita Kumari Kotha, S. Sameen Fatima</i>	
Extraction Based Automatic Text Summarization System with HMM Tagger	421
<i>Suneetha Manne, Zaheer Parvez Shaik Mohd., S. Sameen Fatima</i>	
Recursive Chain Coding Method for Lossless Digital Image Compression	429
<i>T. Meyyappan, S.M. Thamarai, N.M. Jeya Nachiaban</i>	
Short Tandem Repeats in Certain Human Genes Reveal a Positive Correlation towards Evolution	437
<i>Suresh B. Mudunuri, Prudhvi Ravi Raja Reddy Mallidi, Sujan Patnana, S. Pallamsetty, Appa Rao Allam</i>	
Statistical Approach Based Keyword Extraction Aid Dimensionality Reduction	445
<i>M. Ramakrishna Murty, J.V.R. Murthy, P.V.G.D. Prasada Reddy, Suresh Chandra Satapathy</i>	
Hybridization of Rough Set and Differential Evolution Technique for Optimal Features Selection	453
<i>Suresh Chandra Satapathy, Anima Naik</i>	
An Enhanced Scheduling Strategy to Accelerate the Business Performance of the Cloud System	461
<i>T.R. Gopalakrishnan Nair, M. Vaidehi, K.S. Rashmi, V. Suma</i>	
A Novel Approach for Intrusion Detection Using Swarm Intelligence	469
<i>M. Sailaja, R. Kiran Kumar, P. Sita Rama Murty, P.E.S.N. Krishna Prasad</i>	
Genetically Optimized Supplementary Controller for SSSC to Damp Subsynchronous Oscillations	481
<i>Sasmita Padhy, Sidhartha Panda</i>	
Index Page Synthesis Using Genetic Algorithm	489
<i>Ashok Kumar Panda, Satchidananda Dehuri, Isha Padhy</i>	
Web Shield: A Modified Firewall to Detect Malicious Request	497
<i>Urjita Thakar, Omprakash Patel, Lalit Purohit</i>	

Optimal Samples Selection from Gene Expression Microarray Data Using Relational Algebra and Clustering Technique	507
<i>Soumen Kr. Pati, Asit Kr. Das</i>	
Hybrid PSO - Bacterial Foraging Based Intelligent PI Controller Tuning for pH Process	515
<i>G. Petchinathan, G. Saravanakumar, K. Valarmathi, D. Devaraj</i>	
Grid Computing Based NIC Infrastructure: A Step towards IT Enabled India	523
<i>Buddhadeb Pradhan, Rabindra Kumar Shial, Diptendu Sinha Roy</i>	
A Distortion Free Relational Database Watermarking Using Patch Work Method	531
<i>R. Arun, K Praveen, Divya Chandra Bose, Hiran V. Nath</i>	
A High-Speed Two Dimensional Hierarchical Clustering of Microarray Gene Expression Data.	539
<i>R. Priscilla, S. Swamynathan</i>	
Achieving Service Level Agreement in Cloud Environment Using Job Prioritization in Hierarchical Scheduling	547
<i>Rajkumar Rajavel, T. Mala</i>	
Runtime Estimation Aware Scheduling Algorithm for Handling Deadline Based Jobs in Grid Environment	555
<i>Rajkumar Rajavel, T. Mala</i>	
Particle Swarm Optimization Algorithm vs. Genetic Algorithm to Solve Multi-Objective Optimization Problem in Gait Planning of Biped Robot ...	563
<i>Rega Rajendra, Dilip Kumar Pratihar</i>	
Preventing Forest Animals from Train Accidents Using Outlier-Analysis Algorithm in WSN	571
<i>V.P. Jayachitra, Sumalatha Ramachandran</i>	
An Iterative Suffix Stripping Tamil Stemmer	583
<i>Vivek Anandan Ramachandran, Ilango Krishnamurthi</i>	
Combining Classification Algorithm with DOM Algorithm for Web Information Extraction – A Hybrid Approach	591
<i>Venkat Ramana Bhavanasi, A. Damodaram</i>	
Computerized Lesion Detection in Colposcopy Cervix Images Based on Statistical Features Using Bayes Classifier	597
<i>Pazhayanoor Seethapathy RamaPraba, H. Ranganathan</i>	

Heterogeneous Matchmaking Approaches for Semantic Web Service Discovery Using OWL-S	605
<i>P. Ravinder Reddy, A. Damodaram, A.V. Krishna Prasad</i>	
Design and Implementation of Affective E-Learning Strategy Based on Facial Emotion Recognition	613
<i>Arindam Ray, Amlan Chakrabarti</i>	
A Relaxed Parzen Window Based Multifeatured Fuzzy-GIS Model to Forecast Facility Locations (RPWMFGISFFL)	623
<i>Parthajit Roy, Jyotsna Kumar Mandal</i>	
Review on Cost Effective Software Engineering Using Program Slicing Techniques	631
<i>S. Koushik, R. Selvarani</i>	
Binarization of Document Images Using Hierarchical Histogram Equalization Technique with Linearly Merged Membership Function	639
<i>Satadal Saha, Subhadip Basu, Mita Nasipuri</i>	
License Plate Localization Using Vertical Edge Map and Hough Transform Based Technique	649
<i>Satadal Saha, Subhadip Basu, Mita Nasipuri</i>	
Hierarchical Cluster Based Query-Driven Routing Protocol for Wireless Sensor Networks	657
<i>Soumyabrata Saha, Rituparna Chaki</i>	
Application of Dynamic Clustering Using ADE to Transportation Planning	669
<i>Akundi Sai Hanuman, Sesham Anand, A. Vinaya Babu, A. Govardhan</i>	
NLMS Algorithm Based CMA Channel Equalization through an Adaptive MMSE Equalizer	679
<i>Rangisetty Nirmala Devi, Tara Saikumar, K. Kishan Rao</i>	
A Comprehensive Study of Particle Swarm Based Multi-objective Optimization	689
<i>Samanthula Mohankrishna, Divya Maheshwari, P. Satyanarayana, Suresh Chandra Satapathy</i>	
Analysis of Similarity Measures with WordNet Based Text Document Clustering	703
<i>Nadella Sandhya, A. Govardhan</i>	
Application of Particle Swarm Optimization for Combined Environmental and Economic Dispatch of IEEE 30 Bus System Using Fuzzy Logic Technique	715
<i>Sankaramurthy Padmini, Teresa George, Medepalli Sandeep</i>	

An Optimal Design to Schedule the Hydro Power Generation Using Lagrangian Relaxation Method	723
<i>Santhoshkumar Maheswari, C. Vijayalakshmi Seshathri</i>	
Automatic Link Generation for the RDF Dump File: A Minimalistic Approach	731
<i>Arup Sarkar, Ujjal Marjit, Utpal Biswas</i>	
A Font Invariant Character Segmentation Technique for Printed Bangla Word Images	739
<i>Ram Sarkar, Samir Malakar, Nibaran Das, Subhadip Basu, Mahantapas Kundu, Mita Nasipuri</i>	
Generative Process Planning Using Heuristic Artificial Intelligence Technique with CAD Modeling	747
<i>Anand Satchidanandam</i>	
Simultaneous Informative Gene Extraction and Cancer Classification Using ACO-AntMiner and ACO-Random Forests	755
<i>Shimantika Sharma, Shameek Ghosh, Narayanan Anantharaman, Valadi K. Jayaraman</i>	
Role Based Approach for Effective Connections in Backbone of Self Organized Wireless Networks	763
<i>Neeta Shirsat, Pravin Game</i>	
Artificial Neural Network Training Using Differential Evolutionary Algorithm for Classification	769
<i>Tapas Si, Simanta Hazra, N.D. Jana</i>	
A New Unsharp Masking Algorithm for Mammography Using Non-linear Enhancement Function	779
<i>Siddharth, Rohit Gupta, Vikrant Bhateja</i>	
Analysis of Cryptographically Replay Attacks and Its Mitigation Mechanism	787
<i>Arun Kumar Singh, Arun K. Misra</i>	
Comparisons of Three Classifier for Classification of Bamboo Plant	795
<i>Krishna Singh, Surendra Singh</i>	
Secure GKA Using SVD Matrix Decomposition and Kronecker Product ...	803
<i>Reddi Siva Ranjani, D. Lalitha Bhaskar, P.S. Avadhani</i>	
Time-Frequency Domain Techniques for Power System Transients Identification	807
<i>Srikanth Pullabhatla, Prasad Chintakayala</i>	

A Modified Kolmogorov-Smirnov Correlation Based Filter Algorithm for Feature Selection	819
<i>Pakkurthi Srinivasu, P.S. Avadhani, Suresh Chandra Satapathy, Tummala Pradeep</i>	
Software Test Effort Estimation Using Particle Swarm Optimization	827
<i>Prasanta Bhattacharya, Praveen Ranjan Srivastava, Bhanu Prasad</i>	
Prediction of E.coli Protein-Protein Interaction Sites Using Inter-Residue Distances and High-Quality-Index Features	837
<i>Brijesh Kumar Sriwastava, Subhadip Basu, Ujjwal Maulik, Dariusz Plewczynski</i>	
More Secured Text Transmission with Dual Phase Message Morphing Algorithm	845
<i>M. James Stephen, P.V.G.D. Prasad Reddy, Ch. Demudu Naidu, Sampangi Sonali, Ch. Heymaraju</i>	
Removal of False Minutiae with Fuzzy Rules from the Extracted Minutiae of Fingerprint Image	853
<i>M. James Stephen, P.V.G.D. Prasad Reddy, Vadlamani Kartheek, Ch. Suresh, Suresh Chandra Satapathy</i>	
Particle Swarm Optimization Algorithm to Find the Location of Facts Controllers for a Transmission Line	861
<i>S. Harish Kiran, C. Subramani, S.S. Dash, M. Arunbhaskar, M. Jagadeeshkumar</i>	
Uncertain Data Classification Using Rough Set Theory	869
<i>G. Vijay Suresh, E. Venkateswara Reddy, E. Srinivasa Reddy</i>	
Design of Composite Web Service to Obtain Best QoS	879
<i>Urjita Thakar, Abhishek Agrawal</i>	
Classification of Rock Textures	887
<i>Thiagarajan Harinie, Ilangovan Janani Chellam, S.B. Sathya Bama, S. Raju, V. Abhaikumar</i>	
Design and Implementation of an Effective Web Server Log Preprocessing System	897
<i>Saritha Vemulapalli, M. Shashi</i>	
Component Based Resource Allocation in Cloud Computing	907
<i>Sumeet S. Vernekar, Pravin Game</i>	
Interval Evidential Reasoning Algorithm for Requirements Prioritization	915
<i>Persis Voola, A. Vinaya Babu</i>	

Evolutionary Based Secured Coding Technique for Mobile Communication Networks	923
<i>Y.V. Srinivasa Murthy, Suresh Chandra Satapathy, A.A.S. Saranya, K. Sundeep Saradhi</i>	
Research and Application of Dynamic Neural Network Based on Reinforcement Learning	931
<i>Anil Kumar Yadav, Ajay Kumar Sachan</i>	
An Effective Defence Mechanism for Detection of DDoS Attack on Application Layer Based on Hidden Markov Model	943
<i>Suresh Limkar, Rakesh Kumar Jha</i>	
Author Index	951

Optimization of Task Allocation Using Quantum Game Theory with Artificial Intelligence

G.R. Brindha¹, S. Anand², S. Prakash³, and P.M. JoePrathap⁴

¹ ICT Dept, School of Computing, SASTRA University,
Thanjavur, Tamil Nadu, India
grbssk@gmail.com

² Mechatronics Dept, B.Tech-Final Year, SASTRA University,
Thanjavur, Tamil Nadu, India
s.anand90@gmail.com

³ Bangalore, Karnataka
prakashselvakumar@gmail.com

⁴ IT Dept, R M D Engineering College,
Tamil Nadu, India
Joeprathappm@rediffmail.com

Abstract. This paper deals about the implication of quantum game theory with the basis of Artificial Intelligence in real time scenario. Though game theory ideas are basically handled with AI techniques, the radiation of quantum computing gives effective and efficient solutions to the classical games and also in various fields like economics, finance etc., The paper focuses on the following issues: study of quantum strategies in game theory applications and analysis of an application of game theory to solve the real time problem of Task Allocation. The strategies, that we have developed, comprise of different meticulous frame work applied in the field of artificial intelligence.

Keywords: Quantum game theory, Task Allocation, Optimization, Artificial Intelligence.

1 Introduction

Research in quantum computation is looking for the consequences of having information encoding, processing and communication exploit the laws of quantum physics, i.e. the laws which govern the ultimate knowledge that we have, today, of the foreign world of elementary particles, as described by quantum mechanics. The recent developments in game theory have shed new light on real time issues comprising various fields like operations management, human resources and even evolutionary biology. The above scenario provides opportunities to expand the scope of game theory for the quantum world. Quantum games present innovative ways to eliminate dilemmas, to cooperate, to revise equilibria and much more. In classical games discrete set or simple is used while coding the player's strategies whereas in a quantum game they are coded as vectors in a Hilbert space H . Since quantum phenomena is mysterious like quantum world, the usage of quantum applications are

less. But with growing regularity a quantum approach offers more advantages than the classical sets [1], [2], [3], [4], [5]. Through this paper we want to encourage the reader that the study on quantum game theory cannot be ignored because existing hi-tech expansions propose that today or in coming days somebody would take full gain of quantum theory and may use quantum approach to solve many of the practical issues. The paper also presents the basics of quantum game theory with its application and explains the novel usage of quantum game theory for optimizing task allocation.

The most generally developed technique to quantum computation makes use of all four postulates in a clear-cut style. The basic physical carrier of data is a quantum bit which is widely known as qubit, with a 2-dimensional state space -postulate *i*; the state of a n -qubit register exist in a $2n$ -dimensional Hilbert space, the tensor product of n 2-dimensional Hilbert spaces- postulate *iv*. Then by using quantum strategies in conventional organization of classical computation, quantum computations are derived. Quantum computations comprises of three steps in sequence: preparation of the opening state of a quantum register -postulate *iii* can be used for that, possibly with postulate *ii*; next, computation, by using deterministic unitary transformations of the register state -postulate *ii*; and finally, output of a result by probabilistic measurement of all or part of the register -postulate *iii*.

2 Quantum Facts

Except very few laboratory models and special commercial models, there is no existence of general purpose quantum computers. In the normal digital computers, basic unit and processing unit is bit, that represent two states 0 or 1. When very tiny substance is used to characterize a zero or one, quantum mechanics dictates the state and the item are called a qubit, i.e., quantum bits. Multiple qubits are called qubits. The qubit state can be in either pure state 0 or 1 or in a superposition where both be present simultaneously. The trade mark of quantum fact is counterintuitive superposition of both 0 and 1 which has the experimental verification through numerous ways and times. All pure states calculations are done at the end of superposition calculations. i.e., for N qubits of superposition calculation will involve 2^N pure states. There is no matching part for this parallel processing in classical computing.

During the measurement of qubit, in a superposition system, pure state comes for observation. This interface of quantum system with environment is termed as decoherence, so that random pick up of one of the pure state occurs. Since qubits share their quantum states, they can be synchronized even though they are not physically close. This is named as entanglement. When two qubits are entangled and observed if one in 0 state and the other will be in 1 state and vice versa. Till an observation is done, these two qubits are in a superposition. Even both are far away, if one state is known, the other state is also known. Quantum phenomena of superposition, entanglement give way to all new techniques of computing and processing information. Several exciting algorithms have been discovered that take benefit of quantum facts.

3 Mathematical Notations

The pure quantum state of a qubit is represented as follows, $|0\rangle$ represents pure state 0 and $|1\rangle$ represents 1. So an equivalent Qubit is represented by the notation $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ where α and β are complex numbers $|\alpha|^2 + |\beta|^2 = 1$. Here $|\alpha|^2$ probability is 0 and $|\beta|^2$ probability is 1. An n qubit register is represented by

$$|\psi\rangle = \sum_{\mathbf{x} \in \{0, 1\}^n} a_{\mathbf{k}}|\mathbf{k}\rangle, \text{ where } 1 = \sum_{\mathbf{x} \in \{0, 1\}^n} |a_{\mathbf{k}}|^2 \tag{1}$$

This gets 2^n complex numbers (Hilbert space) to totally explain its state. The composite system of 2 qubits, q_1 and q_2 written as $|q_1q_2\rangle$ is

$$|q_1, q_2\rangle = |q_1\rangle \otimes |q_2\rangle = \begin{bmatrix} \alpha_1 \\ \beta_1 \end{bmatrix} \otimes \begin{bmatrix} \alpha_2 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \alpha_1\alpha_2 \\ \alpha_1\beta_2 \\ \beta_1\alpha_2 \\ \beta_1\beta_2 \end{bmatrix} \tag{2}$$

The composite system of 2 qubits is nothing but tensor product of qubits.

4 Quantization of Games

Classical games generally cannot be quantized in a exclusive way because they are only asymptotical shade of a broad range of quantum models. There are two definite amendments of classical simulation games.

- Prequantization: The game becomes a reversal operation on qubits when it is reclassified. This operation represents player's strategies. This strategy is also called quantum coherence.
- Quantization: Decrease the count of qubits and allow random unitary 2 transformations so that the fundamental attributes of the classical game are preserved. Now secondary qubits can be launched so that all quantum subtleties can be possibly explored

The game strategies can influence each other and form collective strategies in quantum games, which is one of the most attractive features. In Fig. 1 measure of

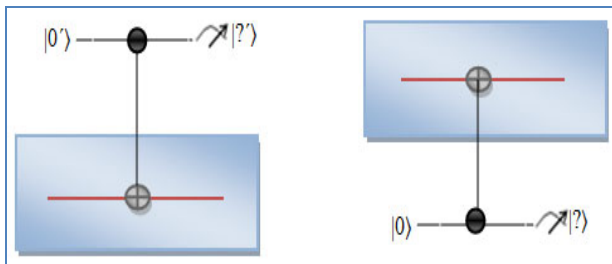


Fig. 1. Determination of others' strategies. The sign at the right ends of lines depicting qubits symbolizes measurement.

observable X' and X is depicted. Entangled quantum strategies will exist only when players in question are unaware of their strategies, since any type of measurement would destroy achievable entanglement of strategies.

5 Recognition of Strategies

The identification if the proper quantization strategy poses as a major problem in the implantation of quantum computation in games. The importance of the proper strategy identification is emphasized by the fact that any quantum computation can be adapted into a quantum game by employing the relevant terminologies that are pertinent to the game theory. The strategies adopted are dependent on the type of the games. Depending upon the information that is presented to the players, there are various classifications of games. One of the most important types of game is the Complete Information Static Games. In this type of games, the players are not allowed to communicate or share information. Since the player cannot obtain any type of information that is pertaining to the other player, the player can only select the available strategies based on the player's own information. Thus, the result that is expected out of this will most probably not be that best possible one. One major assumption that is elemental to the game is that the player selects the strategies that increase their own advantage rather than looking to improve the overall advantage of all the players.

When choosing the proper strategies for the games, there is a lot of emphasis in the difference, between measuring qubits and qubits being measured which is analogous to the computer science terminology kernel and shells. Here the kernel is analogous to the part being measure and the shell is analogous to the measuring part. This distinction is based on the properties of the game rather than any properties of the system's physical representation.

6 Review of Game Theory

The application of quantum strategies into games is complex, in the way that the strategies vary to cater the distinct characteristics of each of the games. But despite the complexities involved in the implementation of the quantum strategies, they prove to be rather efficient when compared to that of the classical strategies. Let us consider a number of examples to illustrate this in detail.

In the bargaining game [6], which is a realistic model of bargaining with a fixed amount of resources, the players get the amount that they bid if the sum of the amounts of bidding is lesser than the total resource that is available for them, else they lose everything. A game which uses the complex quantum strategy has the profit in superposition and the market exchanges are polarized. We can also consider the case of the chinos game [7] in which the game has a number of players and during each turn, a player has to guess the total number of coins in a number of player's hands. After a number of games, the player with the maximum number of right guesses is adjudged the winner. In this game, the full quantum strategy is a winning strategy whereas the partial quantum strategy is not as stable as the classical strategy.

Considering the card game [8],[9] in which the player picks a card from a box which has three cards. One card has dots on both sides, one has circles on both sides and the last has dot on one side and circle on the other. If the player picks the card having the same pattern on both sides then the player wins. This unfair game becomes fair when the classical game is quantified. When we take the gun duel game [10], in which there is two or more gunfighter who shoot at each other and the winner is the person who is last standing in the end. The strategies that are laid out by the players are analogical to the quantum strategy because they are not dependent on the prior outcome since the measurement is not taken till the last. Single round of quantum game is similar to classical version but as the game proceeds the interference plays a bigger role. One more important game is the Monty Hall Problem [11],[12]. This is based on the TV contestant show in which the contestant has to choose a door from three which may have a car or a goat behind it. The host opens one of the doors that the contestant did not choose and offers the contestant opportunity to switch the doors. The contestant should switch the door because the probability that the door has the car increases to $2/3$ then. In the quantum game, the game becomes fair between the contestant and the host. If both of them play the game using quantum strategies, there is no equilibrium in pure strategies but there is in mixed strategies which become similar to a classical game. With entanglement, one quantum player has an advantage over the other classic player. The classical game is the same as the quantum one without entanglement. Thus, from the illustrations the advantages of employing quantum strategies is illustrated.

7 Optimization of Task Allocation

In this paper we have analyzed optimization of Task Allocation supported by the basic game theory of multi player multi choice game using quantization. Task allocation is a significant one in day to day life. We start with 2 members and 2 task then we generalize it to N Tasks. The end results show that if the members work in the quantum world, they can always avoid the worst outcome while in the classical world the worst outcome will always occur with certain probability. The probability of the best result can be much upper than that in the classical world if a diverse value is set for the parameter in the opening state.

We assume that N members are planning to finish a job. There are N tasks with same size which have to be finished to complete the job. Since each of the N members does not know other members choices, so each of them can only choose his task randomly. The payoff for a certain member depends on how many members choose the same task as he/she does. The more members choose the same task, the less payoff of this certain member obtains. If all of them choose the same task, the outcome is the worst because of the two things i.e., possible repetition of same task and in turn delay in completing the work.

The probability of this task allocation is A_{Worst}^Q .

$$A_{\text{Worst}}^Q = N! / N^N \quad (3)$$

The superscript C stands for classical state. On the converse, if they all prefer different tasks, the outcome is the best because there will not be any repetition and delay. The probability of this allocation is A_{Best}^Q .

$$A_{Best}^Q = N! / N^N \tag{4}$$

It is clear that in order to evade the worst outcome, the members will do their best to evade choosing the same task. Since they cannot get information from each other, the worst outcome will occur with certain probability $A_{Worst}^Q = N / N^N$. If we quantize this process, the members can positively avoid the worst outcome by applying quantum strategies that too without knowing what other members choose and the probability to get the best outcome can be much higher than that in classical method if a different value of the parameter in the opening state is set.

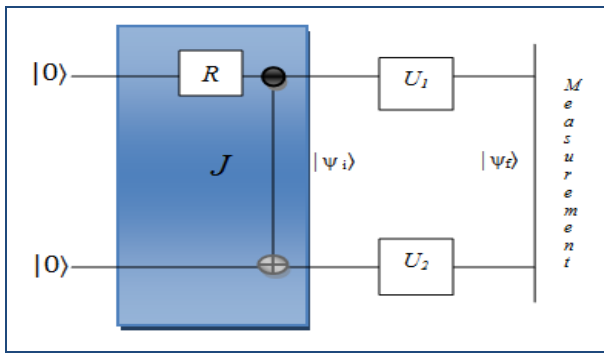


Fig. 2. Task Allocation model

The physical setup of task allocation is shown in the Fig. 2. We send each member a 2-state system in the zero state. The input state is $|\psi\rangle = |00\rangle$. The gate R can be defined as

$$R = 1/\sqrt{2} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \tag{5}$$

Here U_1 and U_2 are the strategies that two members adopt respectively. The State after gate J is $|\psi_i\rangle = 1/\sqrt{2} (|00\rangle - |11\rangle)$. If $U_1 = U_2 = H$, here H is the Hadamard gate, the final state is $|\psi_f\rangle = 1/\sqrt{2} (|01\rangle + |10\rangle)$. From $|\psi_f\rangle$, we can see that the two members are certainly allocated different task. Thus the probability of the best outcome is 1 and the worst situation is avoided, which is the result that members want.

Now the general version of task allocation includes N members and N tasks. The following are initial assumptions: members are numbered from 0 to $N-1$ and the tasks from 0 to $N-1$. If member 0 chooses the road j_0 , member 1 chooses the road j_1, \dots , and member $N-1$ chooses the road j_{N-1} , then the state is described by $|j_0 j_1 \dots j_{N-1}\rangle$. The allocation is started from the state $|00 \dots 0\rangle$. Then a transforming gate is used to obtain the initial state which is denoted by $|\psi_i\rangle$,

$$|\psi_i\rangle = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \omega_N^{k \cdot p} |k \ k \ \dots \ k \rangle \quad (6)$$

Where $\omega_N = e^{2\pi i/N}$. Here p is the parameter that determines the stages of the terms in the expression of the opening state. It is clear that $|\psi_i\rangle$ is symmetric with respect to the swap of the members. Now the members make their decisions on selecting the task. We assume that all of them take up the same strategy. The rationality partially arrives from the allocation symmetry. However the symmetry of the allocation does not promise that all the members prefer the same strategy. Even symmetric games can have asymmetric solutions.[13] While in practice, if a symmetric game has a solution comprise of different strategies, the players would be confused in choosing one from them. Therefore a symmetric solution will be more advantageous than an asymmetric one. Asymmetric solutions will be impossible in symmetric allocations.

It is apparent that U (strategic operator) is an N -dimension unitary operator that executes on the state of an individual member. The precise expression of U is

$$U = (u_{ij})_{N \times N}, u_{ij} = \frac{1}{\sqrt{N}} (\omega_N)^{i \cdot j} \quad (7)$$

Where $i, j = 0, 1, N-1$, and $\omega_N = e^{2\pi i/N}$. U is unitary because

$$\sum_{k=0}^{N-1} u_{ki}^* u_{kj} = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \omega_N^{k(j-i)} = \delta_{ij}, U^\dagger U = I \quad (8)$$

As the members are denoted by U , so $|\psi_i\rangle$ is performed by $U^{\otimes N}$. The final state $|\psi_f\rangle$ is

$$\begin{aligned} |\psi_f\rangle &= U^{\otimes N} |\psi_i\rangle \\ &= \frac{1}{\sqrt{N}} \sum_{k_0=0}^{N-1} \sum_{j_0=0}^{N-1} \dots \sum_{j_{N-1}=0}^{N-1} (\omega_N^{k \cdot p} u_{k j_0} \cdot \dots \cdot u_{k j_{N-1}} |j_0 \cdot \dots \cdot j_{N-1}\rangle) \end{aligned} \quad (9)$$

Therefore the coefficient of $|j_0 \cdot \dots \cdot j_{N-1}\rangle$ is visible

$$\begin{aligned} C_{j_0 \cdot \dots \cdot j_{N-1}} &= \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \omega_N^{k \cdot p} u_{k j_0} \cdot \dots \cdot u_{k j_{N-1}} \\ &= \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \omega_N^{k \cdot m} \end{aligned} \quad (10)$$

Where $m = j_0 + \dots + j_{N-1} + p$.

In some cases members want to guarantee that the payoffs they can at least get is better than the one when they all select the same task, the parameter p can here be set as $p=1$. Thus $|\psi_i\rangle$ is

$$|\psi_i\rangle = 1/\sqrt{N} \sum_{K=0}^{N-1} \omega_N^{k \cdot 1} |k \ k \ \dots \ k \rangle \tag{11}$$

Therefore

$$C_{j_0 \dots j_{N-1}} = (1/\sqrt{N}) \sum_{K=0}^{N+1} \omega_N^k u_{kj_0} \cdot \dots \cdot u_{kj_{N-1}}$$

$$C_{j_0 \dots j_0} = (1/\sqrt{N}) \sum_{K=0}^{N+1} \omega_N^{k \cdot m}$$
(12)

Where $m=j_0 + \dots + j_{N-1} + 1$. When $j_0 = j_1 = \dots = j_{N-1} = j$, the outcome is the worst and coefficient of $|j_0 \dots j_{N-1}\rangle$ is

$$C_{j_0 \dots j_0} = (1/\sqrt{N}) \sum_{K=0}^{N+1} \omega_N^{k \cdot (j_{N+1})} = 0 \tag{13}$$

From equation (13) proved that the worst situation will not occur. The payoffs the got by the members can be definitely better than that of worst outcome.

If the members want to increase the probability of the incident of the best outcome, the parameter can be set as $p=N(N-1)/2$. Thus $|\psi_i\rangle$ is

$$|\psi_i\rangle = 1/\sqrt{N} \sum_{K=0}^{N-1} \omega_N^{k \cdot N(N-1)/2} |k \ k \ \dots \ k \rangle \tag{14}$$

When j_0, j_1, \dots, j_{N-1} are different from each other, there is only one member in one task and the situation is the best. The probability of this case is

$$A_{Best}^Q = N! \cdot |C_{01\dots N-1}|^2 = N \cdot N! / N^N \tag{15}$$

Here the Q represents quantum condition. Compared with the classical probability given in equation (4), we have

$$A_{Best}^Q = N \cdot A_{Best}^C \tag{16}$$

Thus the quantum probability is N times higher than the classical one. From the above explanation, it is clear that the outcome state of the game is closely related to the value of the parameter p in the opening state. By changing the value of parameter p , the needs of members can be fulfilled. If the members expect a higher probability of the best outcome, the value of parameter p can be set as $[N(N-1)]/2$. Also if they demand the eradication of the worst situation, the value of parameter p can be set as 1.

So we suggest that, apart from expected payoff, another payoff is unrepeatable allocation is also important and worth commenting. Assuming that all the members adopt the same strategy, then the game is symmetric with respect to the exchange of the members, the members can always avoid selecting the same way. While in the classical side of this allocation, the worst outcome will happen with certain probability. If the members care most about the probability for the best outcome to occur, in which all members prefer different tasks, the value of the parameter p can be reset to make the probability much higher than that in the classical game.

8 Conclusion

Although a great deal has been achieved in this promising field, there is lots of possibilities still left unexplored in which concepts are likely to be developed by specialists in quantum phenomena. Thus, formulating and studying applications of quantum games within the decision sciences is an important area for future research. The complexity of N-player and N-choice game, is utilized to solve the optimization of task allocation issue, that accompany the worst and the best outcome of the problem. Unlike the classical approach, in the quantum strategy, the solution is derived by setting different values of parameter. With this, members can always meet their requirements by removing the worst out-come or increasing the probability for the best outcome. And this can be accomplished even without knowing the task allocation of other members. Even though quantum game theory is in its insipient stage, there is no doubt that it will become a essential discipline for the up-coming technical society.

References

1. Du, J., Ju, C., Li, H.: Quantum strategy without entanglement. *Journal of Physics A Mathematical and General* 38(7), 1559–1565 (2005)
2. Grabbe, J.O.: An introduction to quantum game theory. Working paper, arxiv: quant-ph/0506219 (2005)
3. Dür, W., Vidal, G., Cirac, J.I.: Three qubits can be entangled in two in- equivalent ways. *Physical Review A* 62 (2000)
4. Eisert, J., Wilkens, M.: Quantum games. *Journal of Modern Optics* 4(14/15), 2543–2556 (2000)
5. Flitney, A.P., Hollenberg, L.C.L.: Nash equilibria in quantum games with generalized two-parameter strategies. *Physics Letters A* 363(5-6), 381–388 (2007)
6. Piotrowski, E.W., Sladkowski, J.: Quantum bargaining games. *Physica A:Statistical Mechanics and its Applications* 308(1-4), 391–401 (2002)
7. Guinea, F., Martin-Delgado, M.A.: Quantum Chinos game: winning strategies through quantum fluctuations. *Journal of Physics A: Mathematical and General* 36(13), 197–204 (2003)
8. Du, J., Ju, C., Li, H.: Quantum strategy without entanglement. *Journal of Physics A Mathematical and General* 38(7), 1559–1565 (2005)
9. Grabbe, J.O.: An introduction to quantum game theory. Working Paper, arxiv: quant-ph/0506219 (2005)

10. Flitney, A.P., Abbott, D.: Quantum two and three person duels. *Journal of Optics B: Quantum and Semiclassical Optics* 6(8), S860–S866 (2004)
11. D’ariano, G.M., Gill, R.D., Keyl, M., Werner, R.F., Kummerer, B., Maassen, H.: The quantum Monty Hall problem. Working Paper, arXiv:quant-ph/0202120 v1 (2002)
12. Flitney, A.P., Abbott, D.: Quantum version of the Monty Hall problem. *Physical Review A* 65(6), 62318 (2002)
13. Du, J.F., Li, H., Xu, X.D., Shi, M.J., Wu, J.H., Zhou, X.Y., Han, R.D.: *Phys. Rev. Lett.* 88, 137902 (2002)

Improving the Grid Scheduling Performance with Fault Tolerance Using Genetic Algorithm

Minu Jacob, Sathya Lakshmi, and Roberts Masilamani

HITS, Chennai

Abstract. In the last few decades we have witnessed the emergence of grid computing as an innovative extension to distributed computing technology, for computing resource sharing among participants in a virtualized collection of organizations. Grid computing entails new challenges as the adaptation of heterogeneous resources unlike homogeneous resources cluster in distributed systems. It is important to maintain proportional fairness in the grid scheduling in order to achieve balanced scheduling. In this paper we propose the importance of genetic algorithm to design schedulers that minimizes the waiting time and maximizes the resource utilization and provides fairness in the grid environment. The resource types and their efficiency are considered in order to maximize their utilization. This paper proposes a solution to maximize the throughput while considering multiple job requests during the scheduling process. The idea of fault tolerance in the crash fault environment will also be implemented based on precautionary method and real time restoration.

1 Introduction

Schedulers are responsible for the management of jobs. Most of the recent literatures treat length of the task as constant [1] which makes the system simple but highly inappropriate in the dynamic environment. The application should have two characteristics (a) Minimize the turn around time (b) Maximizes the resource utilization. We are focused on computational grid and with our scheduler we will be able to use idle cycles effectively. The process of finding the idle cycles are done using optimization techniques and each iteration will give birth to a new set of idle cycles which can be used for the next iteration. Our system reduces the response time without increasing the computational cost because the existing idle cycles of resources are utilized effectively.

2 Background

Related works towards resource scheduling focus on scheduling of a set of independent tasks using various techniques including genetic algorithm [4], data mining [5]. A GA-based scheduler that performs efficiently by minimizing the makespan and flowtime is implemented in [11] which is applicable either in a hierarchical or a simultaneous optimization mode. In all these works the length of the workflow is already known. In our proposed system the length of the workflow is different.

3 Problem Definition

The scheduling algorithms which are used in homogeneous environment can not be compared with that with heterogeneous environment. The scheduling algorithm at this point is based on two factors includes a) The objective function and the b) The specifications and usage. The objective function here is the function the user wants to minimize or maximize. It can be the response time and throughput respectively. The specifications include job requirements, job models and Grid resource models. Authentication, Fault tolerance allocation, and resource reservation also should be closely analyzed and incorporated. Also the consideration of rescheduling and replanning can be considered and can analyze the application components towards single or multiple users as the choice while designing a scheduler. We are considering two types of scheduling problems here: an optimization problem and fault tolerance problem. Two methodologies can be identified for fault tolerance: 1) precautionary and 2) real-time restoration. The idea is to establish a k -connected topology such that every node can reach other nodes over at least k independent paths. Such arrangement will allow the network to seamlessly tolerate the failure of networks. Such arrangement will allow the network to seamlessly tolerate the failure of up to k -Nodes. The provisioning of a high level of connectivity may require the deployment of a large number of resources and may thus be impractical due to their high costs.. When the lost node is a leaf node, no other nodes will be affected. Meanwhile, when the failed node serves as a cut-vertex node in the network, playing the role of a gateway between two sub networks, a serious damage to the network connectivity will be inflicted. Basically with the loss of a cut vertex, the network gets partitioned into disjoint sub networks. This is illustrated in the example of the interactor network in Fig.1. In that Example, the loss of a leaf actor such as A3 will not impact the connectivity of the network. The same applies to no leaf nodes like A12 and A14 when alternate paths are available between the neighbors of the failed actor. Meanwhile, both A1 and A9 are cut vertices, and the failure of either of them will result in two or more disjoint blocks of actors.

3.1 Approach Overview

We present DNRA, a Distributed Node Recovery Algorithm, which opts to efficiently restore the connectivity of an interactor network to its pre-node-failure level.

3.2 The Model

We have developed a model based on Genetic Algorithm where a generation will not begin unless and until the current generation had been finished. The same can be explained with our model where we have a set of tasks, grouped in a workflow. A workflow will not begin next iteration unless the current one is already been finished. We develop a general model to verify the validity of our scheduler. A set of formulae can be developed in order to formulate this. The following parameters can be described for this purpose.

n: numbers of resources
itr: numbers of iterations
N: load for each iteration
N_i: task for each resource
μ_i: instructions per second (MIPS)
T_i: time of iteration
OT: overall time of a workflow

We have chosen the objective function as job completion time and is taken as the minimization of overall completion time. The overall time of workflow can be obtained by finding out the worst case scenario of job-resource allocation and can be formulated as follows

$$OT = \sum_{i=1}^{iter} MAX_i(N_i/\mu_i + Ts_i + Tr_i) \quad (1)$$

The time each iteration can be evaluated based on the following formulae and is obtained by considering the worst case scenario of a task and available resources.

$$T_i = \sum_{k=1}^{tasks} MAX_k(N_i/\mu_i + Ts_i + Tr_i) \quad (2)$$

execution of the task we do not send it to the grid. Thus, (2) can be simplified by (3).

$$T_i \cong MAX_i(N_i/\mu_i) \quad (3)$$

A proportional load is assigned to each following resource to minimize T_i . Thus we do not depend on the resource that offers a worse performance.

$$N_i = N \times \mu_i / \sum_{i=1} \mu_i \quad (4)$$

Where T_{si} is the time that takes sending a task to the resource that is going to execute it and T_{ri} the time that takes receiving the result from a task. We consider the time of sending (T_{si}) and receiving a task (T_{ri}) negligible. Thus, out of this we are able to evaluate the idle cycles of all the resources and thus we can adapt the length of a task according to the total idle cycles. The outcome of this proposal is so simple and efficient. We obtain better load balancing by assigning higher loads to the resources having higher availability. The idea of reduction of the response time and thus the achievement of better throughput is also achieved.

4 Experimental Results

The result of the scheduler is verified based on the simulation techniques. Our scheduler is developed based on Gridsim toolkit [4]. GridSim is a simulation toolkit for application scheduling in parallel and distributed computing. Table 1 shows the resources we have used in our experiment. The MIPS for each processing element is

the maximum MIPS for each resource. This value can change according to each CPU cycle. The delay in transmission is considered as negligible compared to time of execution of task.

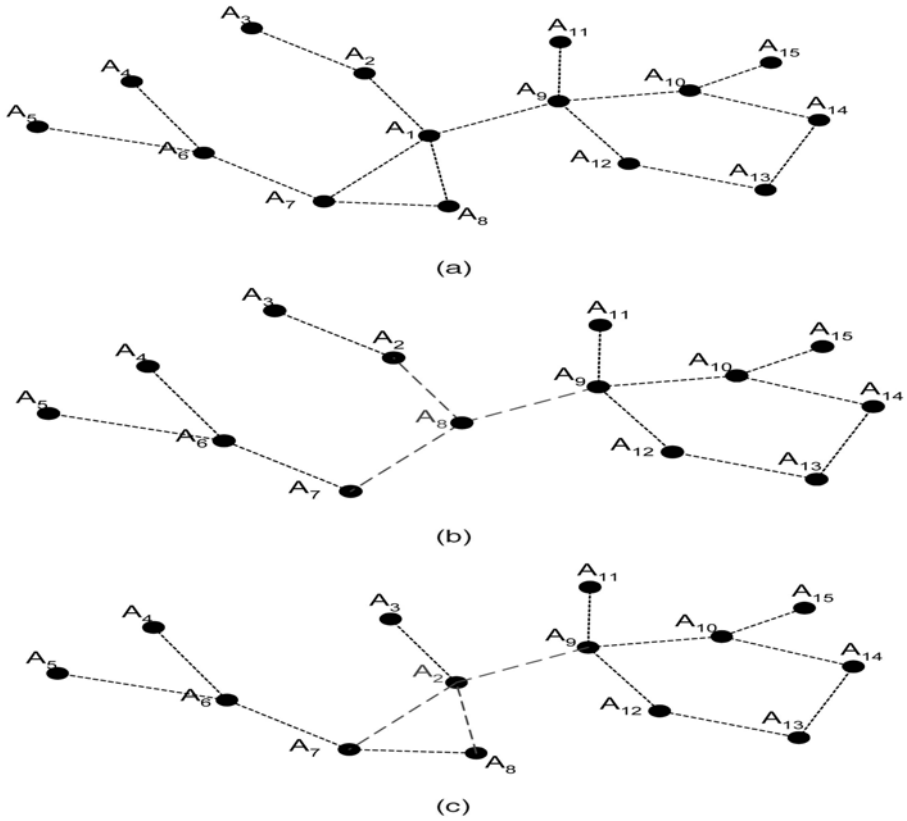


Fig. 1. nteractor Network

Table 1. 500 iterations 50 resources

No.of iteration=500	
Workflow length= 500000000	
No.of resources= 50	
Algorithm	Time(sec.)
FCFS	2637932,84
RR	2740363,3
Balanced	2016518,16

Table 2. Types of resources

Res.	R1	R2	R3	R4	R5
SO	WIN	UNIX	WIN	WIN	WIN
ARQ	IBM	SOLARI S	IBM	IBM	IBM
PE	1	1	1	1	1
MAXMI PS	2000	2500	3000	3500	4000

We compare our proposal with two traditional scheduling algorithms FCFS and Round Robin ((RR) scheduling .We divide our experiments in to two one based on optimization problem where we divide the workflow and consider each as independent tasks. Secondly we describe distributed node recovery algorithm to incorporate fault tolerance.

4.1 Optimization Problem

The GA repeatedly modifies a population of individual solutions. At each step, the GA randomly selects individuals from the current population to be parents and uses them to produce the children for the next generation. Over successive generations, the population evolves toward an optimal solution.

4.1.1 The Experiment

Table 2, 3, 4 shows and fig-2 shows comparison of three schedulers for 50 resources and for 100, 500, 1000 iterations.Fig-2 depicts that higher the number of resources, larger the difference of execution time. That means FCFS and RR regenerates very quickly as generations goes down.

Table 3. 100 iterations 50 resources

No.of iteration=100 Workflow length= 500000000 No.of resources= 50	
Algorithm	Time(sec.)
FCFS	526393,32
RR	548514,52
Balanced	402459,76

Table 4. 1000 iterations 50 resources

No.of iteration=1000 Workflow length= 500000000 No.of resources= 50	
Algorithm	Time(sec.)
FCFS	5269542,88
RR	5487507,1
Balanced	4031486,28

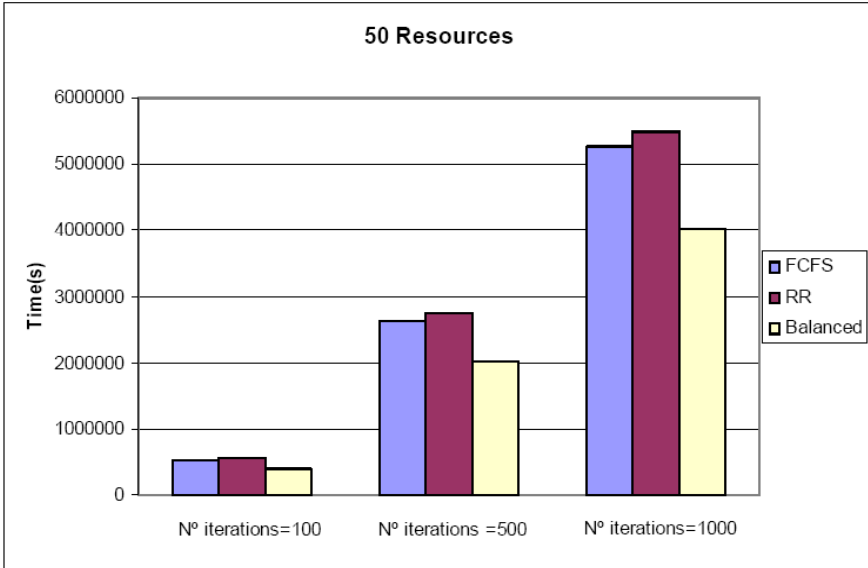


Fig. 2. Experiment with 50 resources and Workflow length= 500000000

4.1.2 Summary of Results

The experiments we have carried out is based on the number of iterations, fixed workflow and number of resources. The experiments are also carried out with various number of resources and found that our balanced scheduler works efficiently as generations goes down. Fig.3,4 and 5 shows 100 ,500,1000 iterations with 25,50,75 and 100 resources. Our balanced scheduler gives shorter execution time for any combination of resources.

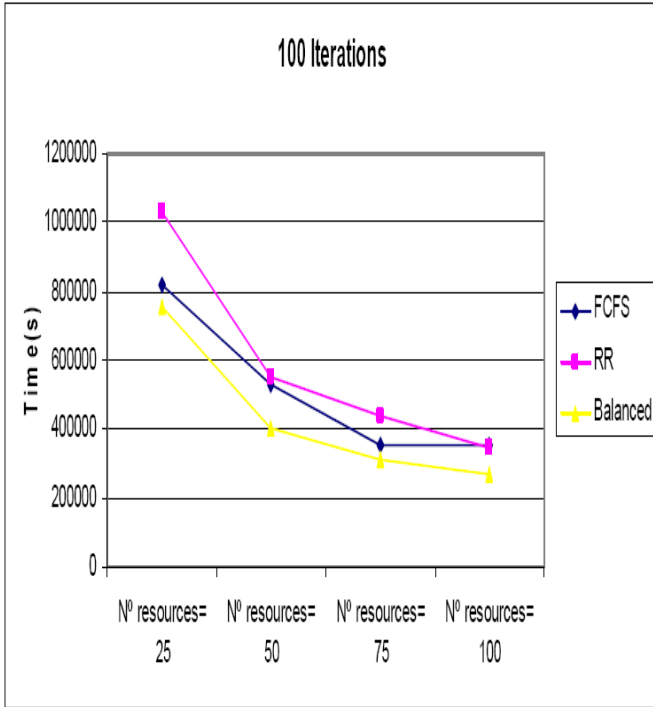


Fig. 3. Experiment with 100 iterations and Workflow length = 500000000

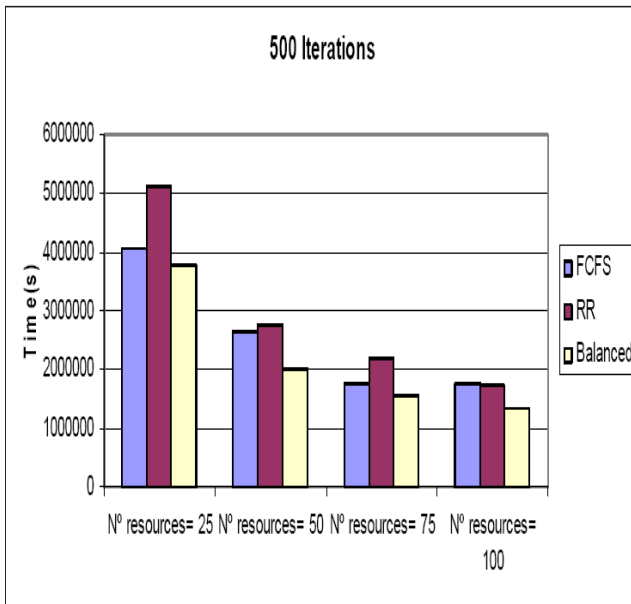


Fig. 4. Experiment with 500 iterations and Workflow length= 500000000

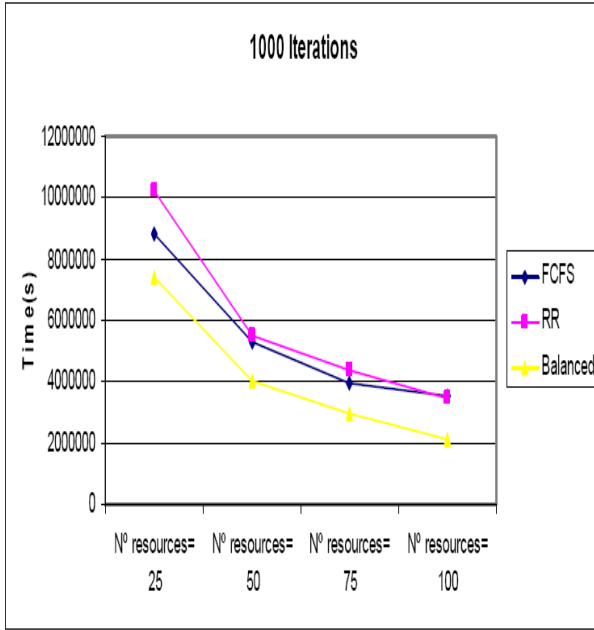


Fig. 5. Experiment with 1000 iterations and Workflow length= 500000000

4.2 Node Recovery

The node recovery algorithm expects nodes A2,A7,A8 and A9 to initiate the recovery process since they are one hop neighbor nodes of A1.The idea of DNRA-1C algorithm is to make either one of these neighbor node to get replaced with the failed node which is A1 in our example. There are some issues in making this replacement a)The prevention of further partitioning b)Which neighbor to replace c) How to ensure the surety of the reconnection of all the nodes by this replacement.DNRA-1C pursues cascaded node relocation in order to sustain connectivity. The selection of node is done based on the degree and proximity of neighbors. Our algorithm prefers the node which is having lowest number of neighbors in order to reduce the overhead. If there is a tie they all will independently come to a conclusion and will assume one as the responsible node for conducting recovery. Among A2, A7, A8, and A9, A9 has the highest node degree and will be thus excluded. Also, A7 has node degree of two, which is larger than that of A2 and A8. Since A2 and A8 have the same node degree and are equidistant to A1,node A8 is picked based on the node ID. Fig. 2b shows the network topology after the recovery. It is worth noting that if A2 is to be picked, A3 may also need to move, as shown in Fig. 2c, and thus, the total travel distance of all involved nodes will be longer than the case when node A8 is selected. Therefore, DNRA-1C may not always yield the optimal results, which is typical for a greedy approach. Nonetheless, as we later discuss, DARA-1C employs a few optimization techniques that proves to be effective in limiting excessive cascaded motion. For

example, in Fig. 2c, if A4 is reachable from A3, we can allow A3 to lose connectivity to A2, and thus, we can further reduce the overhead.

4.2.1 The Algorithm

In this section we analyse the performance of DNRA-1C algorithm by introducing few theorems.

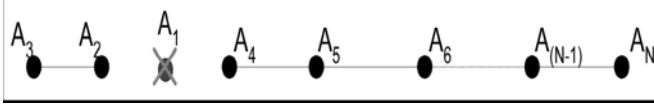


Fig. 6. Worst case scenario

```

DARA-1C(ID, missing_node)
1  repeat
2  BestCandidateID ← FindBestCandidate(TwoHopTable)
3  if ID = BestCandidateID then
4    MoveToLocation(ID, loc(missing_node))
5    exit
6  else
7    Remove BestCandidateID from TwoHopTable
8    pause( 2 * time to travel for distance r)
9  end if
10 until Msg ('RECOVERED') is received

MoveToLocation(ID, newloc)
11 if Neighbor(ID) ≠ NULL then
12   for each j ∈ Neighbor(ID)
13     if IsDependentChild(j) == True then
14       Unicast(j, Msg('MOVING', Siblings(ID)))
// All dependent children will then invoke
// ChildMovOptimizer(j, ID, Siblings(ID))
15   end if
16   end for
17 end if
18 Move(newloc)
19 Broadcast(Msg('RECOVERED'))
20 Update TwoHopTable

ChildMovOptimizer(ID, parent, parentSiblings)
21 for each k ∈ TwoHopTable // meaning ID.TwoHopTable
22   if k ∈ parentSiblings then
23     exit
24   end if
25 end for
26 if ID has not previously moved then
27   DARA(ID, parent)
28 end if

```

Fig. 7. Pseudo code for the distributed node recovery algorithm

5 Conclusion

In this paper we have proposed a scheduling algorithm for grid environment that could be used to do grid scheduling in a fair way. The performance of this algorithm

is proven results in terms of execution time, effectiveness in load balancing and fairness. Utilization of idle cycle also makes it cost effective .We also propose a node recovery mechanism that does fault tolerance in an effective way.

References

- [1] Hwang, R., Gen, M., Katayama, H.: A comparison of multiprocessor task scheduling algorithms with communication costs. *Computers and Operations Research* 35(3), 976–993 (2008)
- [2] Celaya, J., Marchal, L. (eds.): *A Fair Decentralized Scheduler for Bag-of-tasks Applications on Desktop Grids.*- Dept. de Informática e Ingeniería de Sistemas Zaragoza, Spain
- [3] Sulistio, A., Cibej, U., Venugopal, S., Robic, B., Buyya, R.: *A Toolkit for Modelling and Simulating Data Grids: An Extension to GridSim.* In: *Concurrency and Computation: Practice and Experience (CCPE).* Wiley Press, New York (2007)
- [4] Sanchez Santiago, A.J., Yuste, A.J., Munoz Exposito, J.E., Garcia Galan, S., Maqueira Marin, J.M., Bruque, S.: *A dynamic-balanced scheduler for Genetic Algorithms for Grid Computing.* Telecommunication Engineering Department Business Administration and Accounting Department University of Jaén
- [5] Hwang, R., Gen, M., Katayama, H.: A comparison of multiprocessor task scheduling algorithms with communication costs. *Computers and Operations Research* 35(3), 976–993 (2008)
- [6] Duan, R., Prodan, R., Fahringer, T.: *Run-time optimization of grid workflow applications.* In: *7th IEE/ACM International Conference on Grid Computing*, pp. 33–40 (2006)
- [7] Spooner, D.P., Cao, J., Jarvis, S.A., He, L., Nudd, G.R.: *Performance-Aware Workflow Management for Grid Computing.* *The Computer Journal* 48, 347–357 (2005)
- [8] Mandal, A., Kennedy, K., Koelbel, C., Marin, G., Mellor-Crummey, J., Liu, B., Johnsson, L.: *Scheduling strategies for mapping application workflows onto the grid.* In: *Proceedings of the 14th International Symposium on High Performance Distributed Computing (HPDC 2005)*, pp. 125–134 (2005)
- [9] Blythe, J., Jain, S., Deelman, E., Gil, Y., Vahi, K., Mandal, A., Kennedy, K.: *Task scheduling strategies for workflow-based applications in grids.* In: *Proceedings of the Cluster Computing and Grid 2005 (CCGrid 2005)*, pp. 759–767 (2005)
- [10] Weiss, G., Pinedo, M.: *Scheduling Tasks with Exponential Service Times on Non-Identical Processors to Minimize Various Cost Functions.* *Journal of Applied Probability*, 187–202 (1980)
- [11] Carretero, J., Xhafa, F., Abraham, A.: *Genetic based schedulers for grid computing systems*

The Performance Analysis of a Novel Enhanced Artificial Bee Colony Inspired Global Best Harmony Search Algorithm for Clustering

V. Krishnaveni and G. Arumugam

Department of Computer Science,
Madurai Kamaraj University, Madurai, Tamilnadu, India
{krish.h.roban, gurusamyarumugam}@gmail.com

Abstract. Clustering is the unsupervised classification of data items of patterns into groups, each of which should be as homogeneous as possible. The problem of clustering has been addressed in many contexts in many disciplines and this reflects its broad appeal and usefulness in exploratory data analysis. This paper presents a new clustering algorithm, called GHSBEEK which is a combination of the Global best Harmony search (GHS) with features of Artificial Bee Colony (ABC) and K-means algorithms. Global-best Harmony search (GHS) is a derivative-free optimization algorithm, which draws inspiration from the musical process of searching for a perfect state of harmony. It has a remarkable advantage of algorithm simplicity. However, it suffers from a slow search speed. The ABC algorithm is applied to improve the members of the Harmony Memory based on their fitness values and hence improves the convergence rate of the Harmony Search method. The GHSBEEK algorithm has been used for data clustering on several benchmark data sets. The clustering performance of the proposed algorithm is compared with the GHS, PSO, and K-means. The simulation results show that the proposed algorithm outperforms the other algorithms in terms of accuracy, robustness, and convergence speed.

1 Introduction

Cluster analysis is a tool for exploring the structure of data. It is a process in which the objects are grouped into clusters such that the objects from the same clusters are similar and objects from different clusters are dissimilar [1]. Clustering is a challenging job in unsupervised learning which is the process of partitioning a set of objects into an apriori unknown number of clusters while minimizing the within-cluster variability and maximizing the between-cluster variability. Clustering has been used in many engineering and scientific disciplines such as Computer Vision, Information Retrieval, Biology and Market Research [1]. Several clustering algorithm categories have been discussed in the literature, including Hierarchical, Partitional, Density-based and Grid-based algorithms [1]. K-means is one of the popular clustering algorithms. But, K-means algorithm is sensitive to the initial states and always converges to the local optimum solution and hence more stochastic search algorithms are being emerged. In this paper, the GHSBEEK algorithm has been

proposed to overcome this problem as well as to solve the clustering problem. Global-best Harmony search (GHS) is a variation of Harmony search (HS) which is a music-based meta-heuristic optimization algorithm. It was inspired by the observation that the aim of music is to search for a perfect state of harmony. This harmony in music is analogous to find the optimality in an optimization process. It has been proved that GHS outperforms HS when applied to high-dimensional problems [8]. This work shows that the diversity maintenance features of ABC can accelerate the convergence speed of the GHS in the proposed method. Further, the performance of Artificial Bee Colony and Harmony search have been analysed and a novel method for clustering in combination with the K-Means algorithm, called GHSBEEK has been proposed.

In sections 2, 3 and 4, Global-best Harmony Search Algorithm, Artificial Bee Colony Algorithm and K-Means Algorithm have been articulated respectively. In section 5, the GHSBEEK algorithm has been proposed and its efficiency and clustering performance have been analysed using bench mark datasets from UCI repository and finally the paper was concluded in section 6.

2 Global-Best Harmony Search

Inspired by the Particle Swarm Optimization, the GHS algorithm was presented with modified pitch adjustment rule. Unlike the basic HS algorithm, the GHS algorithm generates a new harmony vector X_{new} by making use of the best harmony vector in the Harmony Memory (HM) [8].

Main steps of the algorithm are given below:

- 1: Initialize the problem and algorithm parameters.
- 2: Initialize the harmony memory.
- 3: Improvise a new harmony making use of the best harmony vector
- 4: Update the harmony memory.
- 5: Repeat steps 3-4 until the stopping criterion is met

3 Artificial Bee Colony Algorithm

Artificial Bee Colony (ABC) algorithm [3] for real parameter optimization, is an optimization algorithm which simulates the foraging behaviour of bee colony. The ABC consists of three kinds of bees: employed bees, onlooker bees, and scout bees[3].

In the algorithm, initially, $X_i = (i = 1, \dots, SN)$ solutions are randomly produced in the range of parameters where SN is the number of food sources. In the second step of the algorithm, for each employed bee, whose total number equals to the half of the number of food sources, a new source is produced by (1):

$$V_{ij} = x_{ij} + \emptyset_{ij} (x_{ij} - x_{kj}) \quad (1)$$

where \emptyset_{ij} is a uniformly distributed real random number within the range [-1,1] and k is the index of the solution chosen randomly from the colony. After producing V_{ij} , this

new solution is compared to x_{ij} solution and the employed bee exploits the better source. In the third step of the algorithm, an onlooker bee chooses a food source with the probability (2) and produces a new source in selected food source site by (1). For employed bees, the better source is decided by (2):

$$P_i = \frac{fit_i}{\sum_{j=1}^{SN} fit_j} \quad (2)$$

where fit_i is the fitness of the solution x_i .

After all onlookers are distributed to the sources, sources are checked whether they are to be abandoned. The employed bee associated with the exhausted source becomes a scout and makes a random search in problem domain by (3).

$$x_{ij} = x_j^{min} + (x_j^{max} - x_j^{min}) * rand \quad (3)$$

4 K-Means Algorithm

The goal of data clustering is grouping data into a number of clusters and K -means algorithm is the most popular clustering algorithm. Let $X = (x_1, x_2, \dots, x_N)$ be a set of N data and let each data vector be a p -dimensional vector. Let $C = \{c_1, c_2, \dots, c_k\}$ be a set of K clusters and K denotes the number of cluster centroids which is provided by the user. In K -means algorithm, K cluster centroid vectors are initialized randomly and then each data vector to the class is assigned with the closest centroid vector [8]. In this study, Euclidian metric has been used as a distance metric. The expression is given in (4)

$$D(x_i, c_j) = \sqrt{\sum_{k=1}^P (x_{ik} - c_{jk})^2} \quad (4)$$

After all data are being grouped, the cluster centroid vectors are recalculated using (5)

$$C_j = \frac{1}{n_j} \sum_{x_i \in x_{c_j}} x_i \quad (5)$$

where n_j is the number of data vectors which belong to cluster j . After the above process, the data to the new cluster centroids are reassigned and the process is repeated until a criterion is satisfied. To measure the goodness of the partition, a measure must be defined. A popular performance function for measuring goodness of the partition is the total within-Cluster variance or the total mean-square quantization error (MSE), which is defined in (6).

$$Perf(X, C) = \sum_{i=1}^N \mathbf{Min}\{\|X_i - C_l\|^2 \mid l = 1, \dots, K\} \quad (6)$$

5 The Proposed Algorithm - GHSBEEK

5.1 The Idea Behind

In GHS, the update of the Harmony memory highly depends on the past search experiences. Unfortunately, this inherent shortcoming limits the search ability of the GHS method. The food source exploitation feature of the Artificial Bee Colony method is employed to improve the fitness of the solution candidates in the HM. While the ABC inspired GHS algorithm can be used as a global search strategy across the whole solution space, the K-Means algorithm has been used as a local strategy for improving solutions. The following Pseudo code illustrates the GHSBEEK

- (i) Initialize the Harmony Memory (HM) with initial centroids selected randomly from the original data set
Execute K-Means and calculate fitness for each solution vector
- (ii) Improvise a new Harmony: Define centroids for this solution for $I = 1$ to D do (where D represents Dimension)
 - if (rand > HMCR)
 - begin
 - Randomly select a vector solution from HM
 - Use the food source exploitation feature of ABC mutate the vector by its neighbouring centroid values within limits
 - Execute K-Means and calculate fitness of the mutated solution
 - Compare the fitness values of mutated vector and the randomly selected one
 - Newcentroid [I] = mutated vector if it has better fitness value else the randomly selected one
 - if (rand > PAR)
 - Generate a Newcentroid[I] using the best harmony vector
 - endif
 - end
 - else
 - Newcentroid[I] = Randomly selected vector solution from HM
 - endif
 - Next-for
 - Execute K-Means and Calculate fitness for new harmony
- (iii) Update the harmony memory
- (iv) Check the Stopping Criterion: If the maximum number of improvisations is satisfied, Iteration is terminated else repeat steps (ii) and (iii)
- (v) Select the best Harmony in HM: find the best harmony
Execute K-Means and Calculate fitness for best harmony.
- (vi) Return the best harmony in harmony memory

5.2 Data Clustering and Experimental Setup

The Proposed algorithm has been implemented using MATLAB 7.0 and three data sets were selected from the UCI machine learning repository [12].

For GHS algorithm, parameters were set to the values recommended in [8]. Size of the harmony memory was 15, $HMCR = 0.9$, $PAR = 0.3$, $BW = 0.01$ and the maximum improvisation number was 10000 for all test problems. For the proposed algorithm, the same parameter setting has been maintained. The standard PSO has been used. In this algorithm, the inertia weight ω varies from 0.9 to 0.7 linearly with the iterations and the acceleration factors $c1$ and $c2$ have been kept as 2.0 [6].

The performance evaluation of the proposed GHSBEEK approach for clustering on three different data sets was done and its results were compared with the results of the K-means, PSO, and GHS clustering algorithms.

Motorcycle data ($N = 133$, $d = 2$, $K = 4$): the Motorcycle benchmark consists of a sequence of accelerometer readings through time following a simulated motorcycle crash during an experiment to determine the efficacy of crash helmets.

Iris data ($N = 150$, $d = 4$, $K = 3$): this data set is with 150 random samples of flowers from the iris species *setosa*, *versicolor*, and *virginica*. From each species there are 50 observations for sepal length, sepal width, petal length, and petal width in cm.

Wine data ($N = 178$, $d = 13$, $K = 3$): There are 178 instances with 13 numeric attributes in wine data set. All attributes are continuous. There is no missing attribute value.

For every data set, each algorithm has been applied 30 times individually with random initial solution. Table 1 summarizes the intracluster distances, as defined in (6), obtained from all algorithms for the data sets above. The average, best, and worst solution of fitness from 30 simulations, and standard deviation have been presented in Table 1. Fig. 1, 2 and 3 show the search progress of the average values found by four algorithms over 30 runs for three data sets.

5.3 Experimental Results

From the values in Table 1, it has been concluded that the results obtained by GHSBEEK are clearly better than the other algorithms for all data sets; GHS is a little better than PSO; the K-means is the worst for all data sets.

For Motorcycle data set, the optimum of the fitness function for all algorithms, except K-means, is $2.060e+003$. From the values of the standard deviation, it is observed that the GHSBEEK algorithm is performing better than the other methods. The standard deviation value of GHSBEEK, which is less than 1 represents that the algorithm is converged to the global optimum most of the times.

For Iris data set, GHSBEEK and GHS provide the optimum values and small standard deviation when compared to those of obtained by other methods. The average values of the fitness function for GHSBEEK and GHS are $0.927e+002$ and $0.930e+002$ respectively; the standard deviations for GHSBEEK and HS algorithms are less than 1 which indicates that GHSBEEK and GHS are converged to the global optimum most of the times.

For Wine data set, the results of GHSBEEK algorithm have outperformed the other methods. It has converged to Global optimum most of the times compared to other methods.

Finally, from the graphs shown in Fig. 1, 2 and 3 for all data sets, it has been concluded that GHSBEEK outperforms the other three methods as it converges to the optimal value in a faster manner

Table 1. Comparison of intracluster distances for the four clustering algorithms

Data set	Criteria	GHSBEEK	GHS	PSO	K-means
Motor Cycle	Average	2.078e + 003	2.854e + 003	2.976e + 003	3.412e + 004
	Best	2.070e + 003	2.070e + 003	2.077e + 003	3.187e + 004
	Worst	2.224e + 003	2.934e + 003	3.053e + 003	3.658e + 004
	Std	1.176e + 001	1.198e + 001	1.549e + 001	2.623e + 001
Iris	Average	0.927e + 002	0.930e + 002	0.975e + 002	1.342e + 002
	Best	0.904e + 002	0.904e + 002	0.921e + 002	1.067e + 002
	Worst	0.935e + 002	0.947e + 002	1.053e + 002	1.725e + 002
	Std	1.942e - 001	1.754e + 000	1.7629 + 000	1.736e + 001
Wine	Average	1.652e + 003	1.673e + 003	1.342e + 004	1.642e + 004
	Best	1.603e + 003	1.606e + 003	1.297e + 004	1.607e + 004
	Worst	1.697e + 003	1.698e + 003	1.363e + 004	1.684e + 004
	Std	1.917e - 002	1.146e + 000	1.128e + 001	1.926e + 001

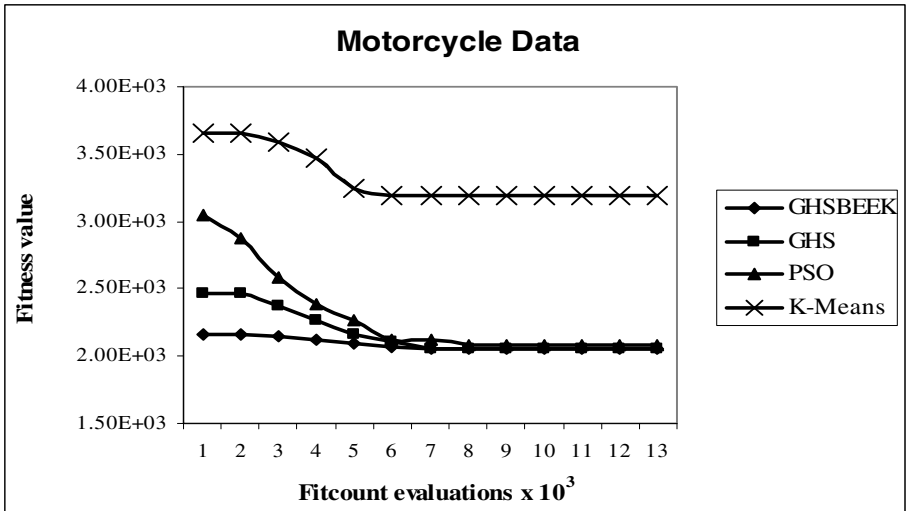


Fig. 1. Comparing the convergence of the proposed GHSBEEK based clustering with other approaches in terms of total Mean-Square quantization Error for Motorcycle data set

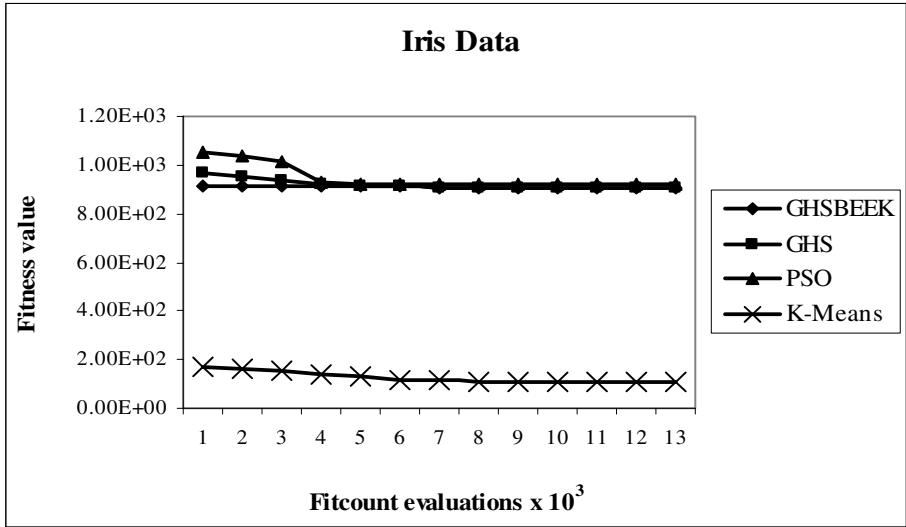


Fig. 2. Comparing the convergence of the proposed GHSBEEK based clustering with other approaches in terms of total Mean-Square quantization Error for Iris data set

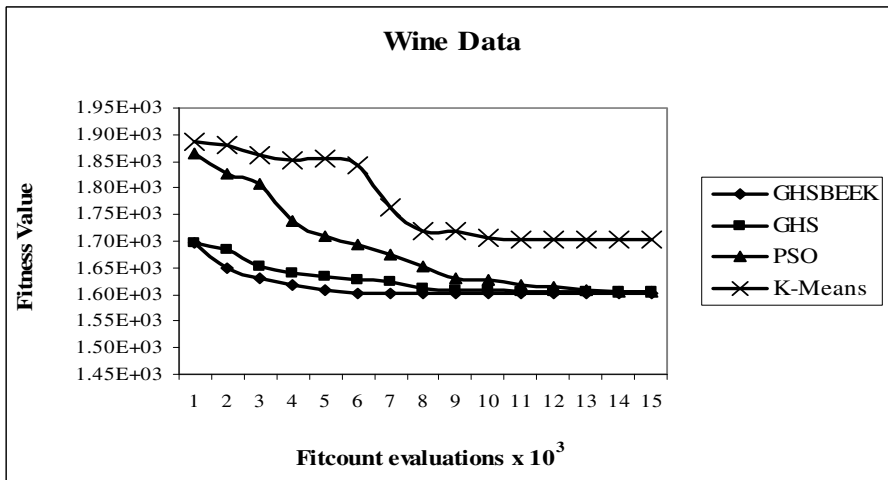


Fig. 3. Comparing the convergence of the proposed GHSBEEK based clustering with other approaches in terms of total Mean-Square quantization Error for Wine data set

6 Conclusion and Future Work

This paper presented a novel algorithm GHSBEEK for solving data clustering problem. The performance of GHS algorithm has been increased by employing the food source exploitation feature of the ABC algorithm which improves the members of the Harmony Memory based on their fitness values and hence improves the

convergence rate of the Global-best Harmony Search method. The exploitation process has been carried out in a controlled way so that the better harmony vectors enjoy the higher selection probability. These actions enabled the speedy update of harmony memory with better solutions and hence caused the search process to move rapidly towards the goal. This enhancement also avoids the problem of getting trapped into the local minima, as the chance of selecting the same harmony vector repeatedly has been minimized. This results in an optimization algorithm which can be used for solving multivariable, multimodal function optimization. This algorithm, in combination with the K-Means clustering algorithm showed significant improvements in the performance in terms of solution quality and convergence speed compared to other optimization algorithms in the data clustering process.

There are many tasks for future work; The GHSBEEK can be applied to real data sets; a metric can be included so that the number of clusters can be found automatically; the other variants of HS such as SGHS and IHS can be used instead of GHS; K-Medoids or Expectation Maximization algorithms can be used instead of K-Means and the results can be compared; finally, the concept of Feature Selection can be included.

References

1. Jain, A.K., Murty, M.N., Flynn, P.J.: Data Clustering: A Review. ACM, 0360-0300/99/0900-0001 (2000)
2. Bratton, D., Kennedy, J.: Defining a Standard for Particle Swarm Optimization. In: Proc. Of the IEEE Swarm Intelligence Symposium (SIS), pp. 120–127 (2007)
3. Karaboga, D., Basturk, B.: Artificial Bee Colony (ABC) Optimization Algorithm for Solving Constrained Optimization Problems. In: Melin, P., Castillo, O., Aguilar, L.T., Kacprzyk, J., Pedrycz, W. (eds.) IFSA 2007. LNCS (LNAI), vol. 4529, pp. 789–798. Springer, Heidelberg (2007)
4. Karaboga, D.: An idea based on honey bee swarm for Numerical optimization. Technical Report TR06, Erciyes University, Engineering faculty, Computer Engineering Department (2005)
5. Lee, K.S., Geem, Z.W.: A new meta-heuristic algorithm for continuous engineering optimization: harmony search theory and practice. *Computer Methods in Applied Mechanics and Engineering* 194(36-38), 3902–3922 (2005)
6. Kennedy, J., Eberhart, R.C.: Particle Swarm Optimization. In: Proc. of the IEEE International Conference on Neural Networks, pp. 1942–1948 (1995)
7. Liu, K., Tan, Y., He, X.: Particle Swarm Optimization Based Learning Method for Process Neural Networks. In: Zhang, L., Lu, B.-L., Kwok, J. (eds.) ISNN 2010, Part I. LNCS, vol. 6063, pp. 280–287. Springer, Heidelberg (2010)
8. Omran, M.G.H., Mahdavi, M.: Global-best Harmony Search. *Appl. Math. Comput.* 198, 643–656 (2008)
9. Redmondand, S.J., Heneghan, C.: A method for initializing the K-means clustering algorithm using kd trees. *Pattern Recognition Letters* 28, 965–973 (2007)
10. Kang, S.L., Geem, Z.W.: A new structural optimization method based on the Harmony search Algorithm. *Computers and Structures* 82(9-10), 781–798 (2004)
11. Geem, Z.W., Kim, J.H., Loganathan, G.V.: Harmony Search Optimization: application to pipe network design. *International Journal of Modeling and Simulation* 22(2), 125–133 (2002)
12. UCI Machine Learning Repository: datasets,
<http://archive.ics.uci.edu/ml/datasets.html>

Artificial Bee Colony Based Image Clustering

Kalyani Manda¹, Suresh Chandra Satapathy², and K. Rajasekhara Rao³

¹ Maharaj Vijayaram Gajapati Raj College of Engineering, Vizianagaram, India
kalyani.mvgr@gmail.com

² ANITS, Vishakapatnam, India
sureshsatapathy@ieee.org

³ KL University, Guntur, India
krr@kluniversity.in

Abstract. The paper presents a novel approach of clustering image datasets with artificial bee colony (ABC) technique. From our simulations it is found that ABC is able to optimize the quality measures of clusters of image datasets. To claim the superiority of ABC based clustering we have compared the outcomes of ABC with the classical K-means and popular Particle Swarm Optimization (PSO) algorithms for the same datasets. The comparisons results reveal the suitability of ABC for image clustering in all image datasets.

Keywords: Image Clustering, K-means, PSO, Artificial bee Colony.

1 Introduction

Image clustering and categorization is a means for high-level description of image content [1]. Image clustering approaches can be broadly categorized to two classes: supervised and unsupervised. All these algorithms further can be classified into two groups: hierarchical and partitional [2][3]. In this paper only partitional algorithm is discussed in which the clustering is formed by minimizing some criteria i.e. squared error function. Hence it can be treated as an optimization problem. The objective here is to minimize the criteria function. K-means[4] is a well known approach for partitional clustering. However, the K-means algorithm is not always able to optimize the mean squared error criterion as it is dependent on initialization values. In this work we have explored the ABC [5] approach for optimizing the mean squared error values. Three benchmark image datasets are chosen for the clustering purpose. In this work we have implemented ABC for clustering. To compare the results obtained with ABC we have simulated K-means and PSO [6] for clustering same datasets. The results reveal that K-means algorithm is trapped in local minima in all the problems whereas PSO and ABC present better results. Compared to PSO, ABC is able to provide more accurate optimized results for all the investigated dataset.

The rest of the paper is organized as follows: An overview of K-means, PSO and ABC image clustering are given in section 2. Section 3 describes the image data set and simulation results. Section 4 concludes the paper, and outlines further improvement.

2 K-Means, PSO, and ABC: Image Clustering Overview

Following are the terminology used to describe the K-means, PSO and DE image clustering algorithms:

- ❖ N_d : number of dimension of image data vector
- ❖ Np : number of image pixels
- ❖ Nc : number of clusters(as provided by the user)
- ❖ Zp : pixel p having N_d dimension
- ❖ m_j : mean of cluster j

To measure the quality of above three algorithms we have chosen a quality metric called as Quantization error (Q_e) and is given by

$$Q_e = \frac{\sum_{j=1}^{N_c} \left[\sum_{\forall z_p \in c_{ij}} \frac{d(z_p, m_j)}{\text{mod}(c_{ij})} \right]}{N_c} \quad (1)$$

$$\text{Where } d(z_p, m_j) = \sqrt{\sum_{k=1}^{N_d} (z_{pk} - m_{jk})^2} \quad (2)$$

and $\text{mod}(c_{ij})$ is the number of data vectors belonging to the cluster j.

We use pixels as the data objects for image clustering. The image is converted into its corresponding RGB values. The gray scale of these values are computed which represents the intensity of the brightness.

2.1 K-Means Algorithm

In K-means algorithm data vectors are grouped into predefined number of clusters. At the beginning the centroids of the predefined clusters are initialized randomly. The dimensions of the centroids are same as the dimension of data vectors. Each pixel is assigned to the cluster based on the closeness, which is determined by the Euclidian distance measure given in equation (2). After all pixels are clustered, the mean of each cluster is recalculated. This process is repeated until no significant changes result for each cluster mean or for some fixed number of iterations.

The K-means algorithm is summarized as

1. Randomly initialize the N_c cluster centroid vectors
2. Repeat
 - a) For each data vector, assign the vector to the class with the closest centroid vector, where the distance to the centroid is determined using equation (2)

b) Recalculate the cluster centroid vectors, using

$$m_j = \frac{1}{n_j} \sum_{\forall z_p \in C_j} z_p$$

until a stopping criterion is satisfied

2.2 PSO Algorithm

Particle swarm optimization (PSO) is a population-based stochastic search process, modeled after the social behavior of a bird flock [6]. The algorithm maintains a population of particles, where each particle represents a potential solution to an optimization problem. In the context of PSO, a swarm refers to a number of potential solutions to the optimization problem, where each potential solution is referred to as a particle. The aim of the PSO is to find the particle position that results in the best evaluation of a given fitness (objective) function.

Each particle represents a position in N_d dimensional space, and is: “flown” through this multi-dimensional search space, adjusting its position toward both

- the particle's best position found thus far. and
- the best position in the neighborhood of that particle.

Each particle i maintains the following information:

- x_i : The **current position** of the particle;
- v_i : The **current velocity**. of the particle;
- y_i : The **personal best position** of the particle.

Using the above notation, a particle's position is adjusted according to

$$v_{i,k}(t+1) = w v_{i,k}(t) + c_1 r_{1,k}(t) (y_{i,k}(t) - x_{i,k}(t)) + c_2 r_{2,k}(t) (\hat{y}_k(t) - x_{i,k}(t)) \quad (3)$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad (4)$$

$$r_{1,j}(t), r_{2,j}(t) \sim U(0,1) \text{ and } k=1, \dots, N_d$$

Where w is the inertia weight, c_1, c_2 are the acceleration constants.

The velocity is thus calculated based on three contributions: (1) a fraction of the previous velocity, (2) the cognitive component which is a function of the distance of the particle from its personal best position, and (3) the social component which is a function of the distance of the particle from the best particle found thus far (i.e. the best of the personal bests) The personal best position of particle i is calculated as

$$y_i(t+1) = \begin{cases} y_i(t) & \text{if } f(x_i(t+1)) \geq f(y_i(t)) \\ x_i(t+1) & \text{if } f(x_i(t+1)) < f(y_i(t)) \end{cases} \quad (5)$$

Where $f()$ is the function evaluation.

Two basic approaches to PSO exist based on the interpretation of the neighborhood of particles [7-8]. Equation (3) reflects the *gbest* version of PSO where, for each particle, the neighborhood is simply the entire swarm. The social component then causes particles to be drawn towards the best particle in the swarm. In the *lbest* PSO model, the swarm is divided into overlapping neighborhoods, and the best particle of each neighborhood is determined. For the *lbest* PSO model, the social component of equation (3) changes to.

$$c_2 r_{2,k} \hat{y}_{j,k}(t) - x_{i,k}(t) \tag{6}$$

where \hat{y}_j is the best particle in the neighborhood of the i^{th} particle.

The PSO is usually executed with repeated application of equations (3) and (4) until a specified number of iterations have been exceeded. Alternatively, the algorithm can be terminated when the velocity updates are close to zero over a number of iterations.

In the context of clustering, a single particle represents the N_c cluster centroid vectors. That is, each particle x_i is constructed as follows:

$$x_i = (m_{i,1}, \dots, m_{ij}, \dots, m_{iN_c}) \tag{7}$$

where m_{ij} refers to the j -th cluster centroid vector of the i -th particle in a cluster. Therefore, a swarm represents a number of candidate clusters for the current data vectors. The fitness of particles is easily measured as the quantization error given in equation (1).

Using the standard *gbest* PSO, data vectors can be clustered as follows:

1. Initialize each particle to contain N_c , randomly selected cluster centroids. For example Iris data set has Four dimension and three clusters. Hence each particle should have 12 i.e 4*3 dimensions. Here n is number of particles and m is the dimension of particles.

X_{11}	X_{12}	X_{13}	-----	X_{1i}	-----	X_{1m}
X_{21}	X_{22}	X_{23}	-----	X_{2i}	-----	X_{2m}
X_{31}	X_{32}	X_{33}	-----	X_{3i}	-----	X_{3m}
			-----		-----	
			-----		-----	
X_{n1}	X_{n2}	X_{n3}	-----	X_{ni}	-----	X_{nm}

2. For $t = 1$ to t_{max} do
 - (a) For each particle i do
 - (b) For each data vector z_p

i) Calculate the Euclidean distance $d(z_p, m_{i,j})$ to all cluster centroids C_{ij}

ii) Assign z_p to cluster C_{ij} such that

$$d(z_p, m_{ij}) = \min_{\forall c=1, \dots, N_c} \{d(z_p, m_{ic})\}$$

iii) Calculate the fitness using equation (1)

(c) Update the global best and local best positions

(d) Update the cluster centroids using equations (3) and (4)

where t_{max} is the maximum number of iterations.

The population-based search of the PSO algorithm reduces the effect that initial conditions have, as opposed to the K-means algorithm; the search starts from multiple positions in parallel.

2.3 Artificial Bee Colony

Artificial Bee Colony (ABC) is one of the most recently defined algorithms by Dervis Karaboga [5] in 2005, motivated by the intelligent behavior of honey bees. The minimal model of forage selection that leads to the emergence of collective intelligence of honey bee swarms consists of three essential components: food sources, employed foragers and unemployed foragers and the model defines two leading modes of the behavior: the recruitment to a nectar source and the abandonment of a source.

i) Food Sources: The value of a food source depends on many factors such as its proximity to the nest, its richness or concentration of its energy, and the ease of extracting this energy. For the sake of simplicity, the “profitability” of a food source can be represented with a single quantity

ii) Employed foragers: They are associated with a particular food source which they are currently exploiting or are “employed” at. They carry with them information about this particular source, its distance and direction from the nest, the profitability of the source and share this information with a certain probability.

iii) Unemployed foragers: They are continually at look out for a food source to exploit. There are two types of unemployed foragers: scouts, searching the environment surrounding the nest for new food sources and onlookers waiting in the nest and establishing a food source through the information shared by employed foragers. The mean number of scouts averaged over conditions is about 5-10%. The exchange of information among bees is the most important occurrence in the formation of the collective knowledge. While examining the entire hive it is possible to distinguish between some parts that commonly exist in all hives. The most important part of the hive with respect to exchanging information is the dancing area. Communication among bees related to the quality of food sources takes place in the dancing area. This dance is called a waggle dance. Since information about all the current rich sources is available to an onlooker on the dance floor, probably she can watch numerous dances and decides to employ herself at the most profitable source. There is a greater probability of onlookers choosing more profitable sources since more information is circulated about the more profitable sources. Employed foragers

share their information with a probability proportional to the profitability of the food source, and the sharing of this information through waggle dancing is longer in duration. Hence, the recruitment is proportional to the profitability of the food source.

The workings of honey bees are explained in this paragraph. Assume that there are two discovered food sources: A and B. At the very beginning, a potential forager will start as unemployed forager. That bee will have no knowledge about the food sources around the nest. There are two possible options for such a bee: (i) It can be a scout and starts searching around the nest spontaneously for a food due to some internal motivation or possible external clue, or (ii) it can be a recruit after watching the waggle dances and starts searching for a food source. After locating the food source, the bee utilizes its own capability to memorize the location and then immediately starts exploiting it. Hence, the bee will become an “employed forager”. The foraging bee takes a load of nectar from the source and returns to the hive and unloads the nectar to a food store. After unloading the food, the bee has the following three options: (i) It becomes an uncommitted follower after abandoning the food source, or (ii) It dances and then recruits nest mates before returning to the same food source, or (iii) It continues to forage at the food source without recruiting other bees.

It is important to note that not all bees start foraging simultaneously. The experiments confirmed that new bees begin foraging at a rate proportional to the difference between the eventual total number of bees and the number of present foraging. The behavior of honeybee foraging for nectar In the case of honey bees, the basic properties on which self organization relies are (i) Positive feedback: As the nectar amount of food sources increases, the number of onlookers visiting them increases, too, (ii) Negative feedback: The exploration process of a food source abandoned by bees is stopped, (iii) Fluctuations: The scouts carry out a random search process for discovering new food sources, and (iv) Multiple interactions: Bees share their information about food source positions with their nest mates on the dance area.

2.4 Suitability of ABC for Clustering

Apart from being used in function optimization [9] ABC can also be used to cluster datasets. The procedure followed is similar to that used in function optimizations. The functions that need to be optimized are the intracluster distances, between the objects belonging to the same cluster and intercluster distances, between objects of different clusters. These intracluster distances measure the compactness of a cluster in ABC while the intercluster distances measure the degree of separation.

The ABC based clustering techniques can also be single objective as well as multi objective. In this project we have taken single objective clustering technique and the optimization function is intracluster distance. However, ABC has too many parameters to adjust. One version needs to be fine tuned for other datasets, so that it can work well in a wide variety of applications. ABC is being used for approaches that can be used across a wide range of applications, as well as for specific applications focused on a specific requirement. There is a lot of scope for research under this algorithm, as it is a very recent technique developed in the year 2005.

3 Images and Simulation Results

The three image clustering algorithms namely K-means, PSO, and ABC have been applied to three types of imagery data, namely MRI brain, Lena, and Mandrill. These data sets have been selected for testing and comparing above three algorithms. The three images chosen comprises of 250x250 8-bit gray scale pixels. The figure 1, figure 5, and figure 9 are the original images of MRI brain, Lena, and Mandrill respectively. A total no. of clusters of 8, 8, and 6 were randomly chosen respectively for MRI brain, Lena, and Mandrill images. The performances of three chosen algorithms are computed by the quantization error given in equation (1) and the intra and inter cluster distances as in [10].

The clustered images of MRI brain, Lena, Mandrill using K-means are shown in figure 2, figure 6, and figure 10 respectively with the quantization error and inter cluster & intra cluster measures shown in the Table 1. For running the PSO we have chosen parameters swarm size as 10, maximum no. of iterations are 30, c_1 & c_2 are 1.042 both equal [8]. The w value is varied as per [8] in every iteration.

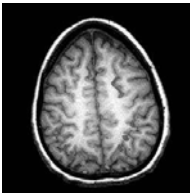


Fig. 1.

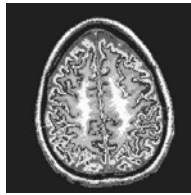


Fig. 2.

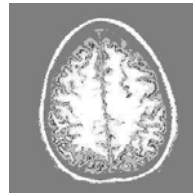


Fig. 3.

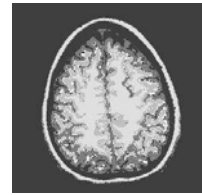


Fig. 4.



Fig. 5.



Fig. 6.



Fig. 7.



Fig. 8.

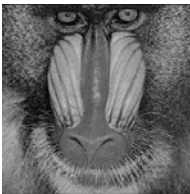


Fig. 9.

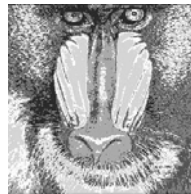


Fig. 10.



Fig. 11.



Fig. 12.

Table 1.

Image	No. of clusters(User chosen)	Algorithm	Quantization error (Q_e)	Intra-cluster distance	Inter-cluster distance
MRI Brain	8	k-means	0.13819	0.18285	30.6242
		PSO	0.13327	0.12553	31.8400
		ABC	0.10012	0.10001	35.4646
Lena	8	k-means	0.07748	0.11933	14.4542
		PSO	0.074451	0.10290	16.2662
		ABC	0.05335	0.0723	19.1212
Mandrill	6	k-means	0.085077	0.13067	23.3382
		PSO	0.083897	0.13844	25.4293
		ABC	0.06985	0.100114	32.4243

In our experiment w_1 (initial weight), w_2 (final weight), v_{\max} are chosen to be 0.9, 0.4 and 10 respectively for best results. The clustered images for PSO MRI brain is shown in figure 3. It can be seen that K-means trapped in local optimum and could not classify the clusters correctly. PSO in other hand is not trapped in this local minimum. This can be verified from the quantization error measure given in the Table 1. The quantization error is **0.10012** which is less than the value for K-means. The results shown in the table 1 clearly indicates the superiority of ABC over other two approaches such as K-means and PSO. In all datasets the quality measures like quantization error (Q_e), intra and inter cluster distances are found to be better for ABC over other two algorithms.

4 Conclusion

This paper presented a novel approach of clustering image dataset with ABC. The ABC clustering results are compared with well known K-means and PSO clustering for all investigated dataset. It was shown that PSO and ABC produced better result compared to K-means with respect to the quantization error, inter- and intra-cluster distances. The local optima problem of K-means was alleviated using PSO and ABC further improved the results.

References

1. Goldberger, J., Gordon, S., Greenspan: Unsupervised image-set clustering using an information theoretic framework. IEEE Trans. on Image Processing 15(2), 449–458 (2006)
2. Frigui, H., Krishnapuram, R.: A Robust Competitive Clustering Algorithm with Applications in Computer Vision. IEEE Transactions on Pattern Analysis and Machine Intelligence 21(5) (1999)

3. Leung, Y., Zhang, J., Xu, Z.: Clustering by Space-Space filtering. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 22(12), 1396–1410 (2000)
4. Tou, J.T., Gonzalez, R.C.: *Pattern Recognition Principles*. Addison-Wesley, Reading (1974)
5. Basturk, B., Karaboga, D.: An Artificial Bee Colony (ABC) Algorithm for Numeric function Optimization. In: *IEEE Swarm Intelligence Symposium 2006*, Indianapolis, Indiana, USA, May 12-14 (2006)
6. Kennedy, J., Eberhart, R.C.: Particle Swarm Optimization. In: *Proceedings of the IEEE International Joint Conference on Neural Networks*, vol. 4, pp. 1942–1948 (1995)
7. Kennedy, J., Mendes, R.: Population Structure and Particle Swarm Performance. In: *International Proceedings of the 2002 Congress on Evolutionary Computation*, pp. 1671–1675. IEEE service Center, Piscataway (2002)
8. Suganthan, P.N.: Particle Swarm Optimizer with Neighborhood Operator. In: *Proc. Congress On Evolutionary Computation*, Washington D.C, USA, pp. 1958–1961. IEEE Service Center, Piscataway (1999)
9. Karaboga, D.: An idea based on Honey Bee Swarm for Numerical Otimization: Technical report-TR06, Erciyes University, Engineering Faculty, Computer Engineering Department (2005)
10. Omran, M., Engelbrecht, A.P., Salman, A.: Image Classification using Particle Swarm Optimization. In: *Proceedings of the 4th Asia-Pacific Conference on Simultaed Evolution and Learning*, Singapore (2002)

Novel Approach to Polygamous Selection in Genetic Algorithms

Rakesh Kumar¹ and Jyotishree²

¹ DCSA, Kurukshetra University, Kurukshetra, India
rsagwal@gmail.com

² DCSA, Guru Nanak Girls College, Yamuna Nagar, India
jyotishreer@gmail.com

Abstract. Genetic algorithms are adaptive heuristic search algorithms which have been successfully used in a number of applications and their performance are mainly influenced by selection operator. In this paper three variants of polygamous selection, a special case of elitism where the best individual of the population act as one parent for mating with other chromosomes in all crossover operations, are proposed, their performances are compared along with other selection approaches such as roulette wheel, rank, annealed etc.

Keywords: Elitism, Genetic Algorithms, Polygamy, Selection.

1 Introduction

Genetic algorithms are random search algorithms that were invented by John Holland in 1975 [1]. They follow the genetic process of biological evolution. They were defined as adaptive heuristic search algorithms based on the evolutionary ideas of natural selection and natural genetics by David Goldberg [2]. A typical genetic algorithm is composed of three main operators – Selection, Crossover and Mutation. A genetic algorithm is an iterative procedure which operates on population of constant size where each individual has a specific fitness value depending on the objective function. Individuals from current generation are selected according to their fitness value and produce offsprings using crossover to form the next generation of individuals. Mutation operator maintains diversity in population. Genetic algorithms are based on Darwin's principle of "Survival of Fittest", so better fit individuals are carried forward to next generation leaving behind the less fit ones [2]. The process of forming next generation of individuals by replacing or removing some offsprings or parent individuals is done by replacement operator. Genetic algorithms iterates till maximum number of generations is reached or until the optimal solution is achieved.

Basic Genetic Algorithm:

```
Procedure GA(tourlength,  $\theta$ , n, r, m, ngen)
//tourlength function evaluates individuals in population
// $\theta$  is the fitness threshold to determine when to halt
//In TSP, best fitness is minimum value of tourlength
// n is the population size in each generation (say 100)
```

```

// r is crossover probability(0.7)& m is mutation rate (0.01)
// ngen is total number of generations
P := generate n individuals at random
// h represents the chromosome in the population P
i:=1
while (min(tourlength(hi)) <  $\theta$  or i <= ngen do
{ //Reproduction step:
  //Select n/2 individuals of P as per any selection method
  Call Select(P,n,r) and store in L
  Probabilistically select (1-r) n individuals from L
  and store in k1 and k2 //mating pool for crossover
  foreach pair selected (k1,k2), produce two offspring
  by applying PMX crossover operator and add to S
  //mutation step
  Choose m% of S and Mutate chromosomes by inversion
  P := S //next generation depending on replacement
  i:=i+1
}
Find b such that tourlength(b) = min(tourlength(hi))
return(b)
end proc.

```

In this paper, a different selection approach – polygamy is considered. Polygamy refers to special case of elitism where the best individual of the population is treated as one parent for mating with other chromosomes in all crossover operations. The paper proposes three different approaches of polygamy with an aim to choose the best chromosome so as to retain good characteristics in the new generation and compares the performance of genetic algorithm in different cases proposed. The paper is organized in the following sections. Research work related to polygamy and replacement strategies have been reviewed in section 2. Different notations used throughout the paper are given in section 3. Algorithms of different approaches for selection and polygamy are described in section 4. These approaches are implemented on Travelling Salesman Problem to test. Implementation procedure and computational results are provided in section 5 and concluding remarks are given in section 6.

2 Related Work

Polygamy is a mating system in which a single individual of one gender mates with several individuals of opposite gender. Polygamy has two forms – Polygyny and Polyandry. In Polygyny, one male individual mates with several females of the respective species as in elk, fur seals etc. In Polyandry, one female mates with more than one male during a breeding season like Honey bees [3]. An improved genetic algorithm based on polygyny was proposed by Gu Min and Yang Feng wherein the population had one father, many mothers and some bachelors. Father and mothers mated with each other using crossover and bachelors participated only in mutation operation [4]. Al jaddan et al. applied different selection operators on eight test function and compared the performance of genetic algorithm in terms of various

criteria like convergence, time, and reliability [5]. Generational and Steady state are two forms of replacement. In generational replacement, entire population of genomes is replaced at each generation. In this case, generations are non-overlapping. In steady state replacement, new individuals are inserted in the population as soon as they are created [6,7]. The $(\mu+1)$ approach was the first steady state replacement strategy introduced by Rechenberg in 1973 and had parent population greater than one ($\mu > 1$) [8]. De Jong introduced the generation gap G as a parameter to genetic algorithm where a percentage of population is chosen via fitness proportionate selection to undergo crossover and mutation[9]. Schwefel proposed $(\mu+\lambda)$ and (μ,λ) models that correspond to overlapping and non-overlapping populations [10].

3 Notations and Definitions

Some of the symbols and notations used in the paper are listed below:

Symbol	Meaning	Symbol	Meaning
ngen	Total number of generations	F_{best}	Best fitness value
nogen	Current number of generation	F_{avg}	Average fitness value
N	Total population size		
RWS	Roulette Wheel selection with generational replacement	RS	Rank selection with generational replacement
AS	Annealed selection with generational replacement	$F_{i,j}$	Fitness of jth individual in ith generation
mpool	Number of chromosomes in mating pool	$FX_{i,j}$	Fitness of individual in Annealed selection

4 Various Approaches for Selection, Polygamy and Replacement

Selection operation is used to choose the best fit individuals from the population for crossover operation. Selection of individuals in the population is fitness dependent and is done using different algorithms [11]. Selection chooses more fit individuals in analogy to Darwin's theory of evolution – survival of fittest [12]. There are many methods in selecting the best chromosomes such as roulette wheel selection, rank selection etc. Replacement operator chooses the offsprings that will stay in the population and the individuals that would be replaced to form the next generation. Polygamy is special case of selection and has biological evidences in natural evolution. In this case, the best fit individual in the current generation would act as one parent in all the crossover operations to create the next generation. The paper analyses the comparison of roulette wheel selection, rank selection and annealed selection [13] and effect of these selection operators in combination with polygamy, $\mu+\lambda$ polygamy and extended $\mu+\lambda$ polygamy.

4.1 Roulette Wheel Selection

Roulette wheel selection technique places all the individuals in the population on virtual roulette wheel according to their fitness value [2,9,11]. Roulette wheel

selection uses exploitation technique and individuals with higher fitness have more probability of selection.

4.2 Rank Selection

Rank Selection sorts the population first according to fitness value and ranks them. Then every individual is allocated selection probability with respect to its rank [14]. Individuals are selected as per their selection probability. It is exploratory in nature.

4.3 Annealed Selection

The annealed selection approach is to blend the exploratory and exploitive nature of rank selection and roulette wheel selection respectively. The perfect blend of the two approaches is achieved by computing fitness value of each individual as per the current generation number as under:

$$FX_{i,j} = F_{i,j} / ((ngen+1) - nogen) \quad (4)$$

Selection pressure is changed with changing generation number [13].

Algorithm for annealed selection is:

```
Annealed selection
  Set l=1, j=1, i=nogen
  While j<=N
  {
     $FX_{i,j} = F_{i,j} / ((ngen+1) - nogen)$ 
  }
  Set j=1, S=0
  While j<=N
  {
     $S=S+FX_{i,j}$ 
  }
  While l <= mpool
  {
    Generate random number  $r$  from interval (0,S)
    Set j=1, S=0
    While j<=N
    {
       $C_j=C_{j-1}+FX_{i,j}$ 
      If  $r<=C_j$ , Select the individual j
    }
    l=l+1
  }
}
```

4.4 Polygamy

Polygamy is special kind of selection which has biological evidences in nature as in the case of honey bee, lion, leech etc. This approach is based on the biological fact that selecting the most fit parent would lead to fitter offsprings for the next generation [3]. Salient Features of Polygamous selection are:

- The best fit individual of the population is selected as one parent and will participate in all crossover operations.
- Second parent is selected using any of the three selections discussed earlier.
- The best parent selected for polygamy participates in crossover in its respective generation only.

- Next generation of population is generated using generational replacement.

Module selecting the best parent is as follows:

```
Polygamy (P, n)
    Select  $h_i$  having min(tourlength) and store in  $k_1$ 
    Call Select(P, n, r) and store in L
End.
```

Outline of Genetic algorithm implementing polygamy is given below:

```
Procedure GA(tourlength,  $\theta$ , n, r, m, ngen)
    :
    //Reproduction step:
    Call Polygamy(P, n)
    :
end proc.
```

4.5 $\mu+\lambda$ Polygamy

$\mu+\lambda$ polygamy is combination of polygamy and competitive elitism. Salient features of $\mu+\lambda$ polygamy are:

- The best individual from the pool of current and the previous generation is selected as one parent that will participate in all crossover operations.
- The second parent is selected depending on earlier discussed selection methods.
- The best parent selected can participate in crossover in successive generations.
- Offsprings generated follow $\mu+\lambda$ replacement strategy to form the next generation.

Genetic algorithm implementing $\mu+\lambda$ polygamy is given below:

```
Procedure GA(tourlength,  $\theta$ , n, r, m, ngen)
    :
    //Reproduction step:
    Call Polygamy(P, n)
    :
    pb:=min(tourlength( $h_i$ ))
end proc.
```

4.6 Extended $\mu+\lambda$ Polygamy

Polygamous selection leads to premature convergence in certain cases. This may be due to loss of diversity by repeated selection of same best parent in each generation. Extended $\mu+\lambda$ polygamy suggests a novel idea of polygamy by limiting the best parent to participate in crossover in consecutive generations. Its salient features are:

- The best individual from the pool of consecutive generations is selected as one parent that will participate in all crossover operations.
- If the best parent selected is same as that of last crossover, then it is replaced by second best individual in the respective generation.

- Second parent for crossover is selected using any of the above selection methods.
- Best parent selected cannot participate in crossover in consecutive generations.
- Offsprings that form the next generation follow $\mu+\lambda$ replacement strategy.

Genetic algorithm implementing extended $\mu+\lambda$ polygamy is given below:

```

Procedure GA(tourlength,  $\theta$ , n, r, m, ngen)
:
  Call Polygamy(P, n)
  If pb= $k_1$ , replace  $k_1$  by next  $h_i$  having min(tourlength)
:
  pb:=min(tourlength( $h_i$ ))
end proc.

```

5 Implementation and Observation

In this paper, code for genetic algorithm is developed using MATLAB for benchmark TSP using Eil51 population as test problem. The code was run for 100 generations using same parameters in different cases of selection and performance was compared in terms of minimum tour length (F_{best}) and average tour length (F_{avg}). Table 1 lists the data for F_{best} and F_{avg} for Eil 51 population in different approaches of selection. Fig. 1 depicts the comparison of average tour length F_{avg} and Fig. 2 depicts the comparison of minimum tour length F_{best} in twelve different cases.

Table 1. Comparison of Different Approaches for Eil 51 population

Method	Favg	Fbest
RWS	1720.973	1371.3383
RS	1667.4847	1387.9336
AS	1515.5429	1315.1413
Polygamy +RWS	1230.7096	1214.776
Polygamy +RS	1106.0641	1073.5741
Polygamy +AS	1190.3857	1171.1816
$\mu+\lambda$ Polygamy +RWS	1201.0688	1183.7948
$\mu+\lambda$ Polygamy +RS	1128.3626	1094.5386
$\mu+\lambda$ Polygamy +AS	1300.7513	1288.4802
Extended $\mu+\lambda$ Polygamy +RWS	1198.0254	1138.488
Extended $\mu+\lambda$ Polygamy +RS	1442.51	1284.19
Extended $\mu+\lambda$ Polygamy +AS	1154.1404	1132.5966

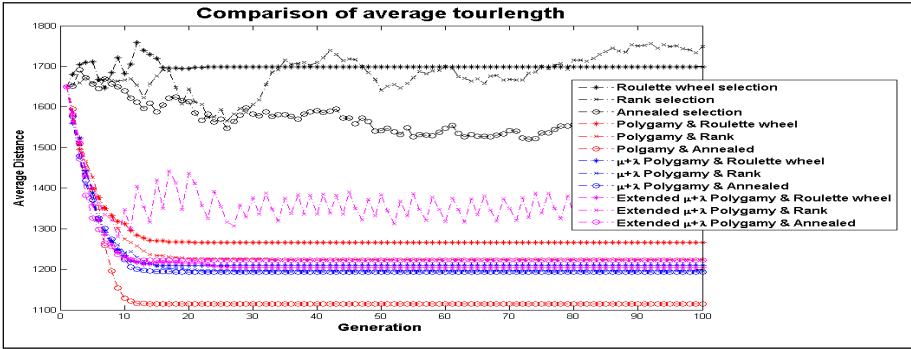


Fig. 1. Average Tour Length vs. Generation

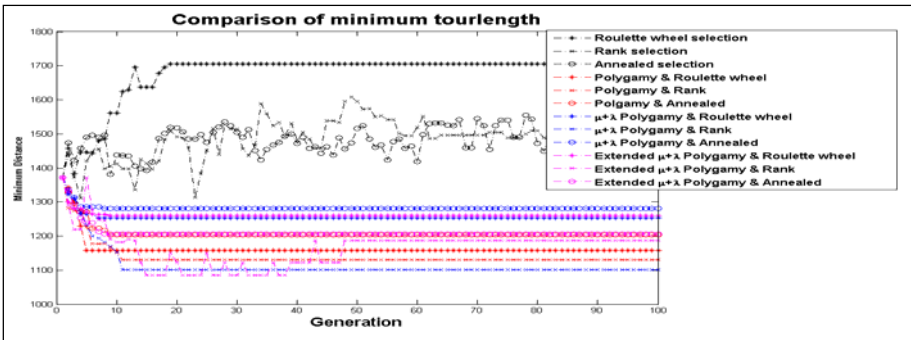


Fig. 2. Minimum Tour Length vs. Generation

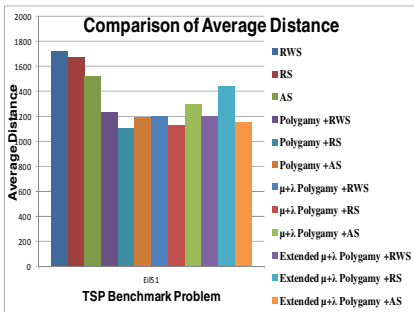


Fig. 3. Comparison of Average tour length

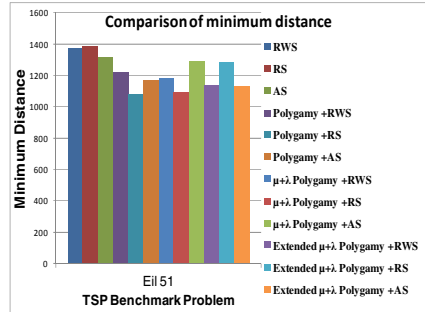


Fig. 4. Comparison of Minimum tour length

It was observed that among the three selection approaches considered, the annealed selection is more promising. The results improved drastically on introduction of polygamy. It is very much clear from the graphs Polygamy is better than simple selection. On further experimentation with $\mu+\lambda$ and Extended $\mu+\lambda$ polygamy, it was observed that the results improved and were even better than or at par with polygamy.

6 Conclusion

The paper compared three different approaches for polygamy using different replacement strategies. It was found that polygamy resulted in better results and detailed analysis suggested that polygamy with generational update led to early convergence due to lack of diversity. Further, the two modified polygamy techniques were compared using $\mu+\lambda$ replacement and resulted in better performance than generational update. $\mu+\lambda$ replacement with polygamy has its biological evidence in case of lions. Extended $\mu+\lambda$ polygamy maintained diversity in population and gave better results. Seeing this result, it can be thought of to have varying dying periods for the best parent in polygamy. This may lead to introduction of diversity in population and would delay or avoid premature convergence.

References

1. Holland, J.: *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor (1975) ISBN 0262581116
2. Goldberg, D.E.: *Genetic Algorithms in Search, Optimisation and Machine Learning*. Addison Wesley Longman, Inc. (1989) ISBN 0-201-15767-5
3. Paxton, R.J.: Male Mating Behaviour and Mating Systems of Bees: an Overview. *Apidologie* 36, 145–156 (2005), Article published by EDP Sciences, <http://dx.doi.org/10.1051/apido:2005007>
4. Gu, M., Yang, F.: An Improved Genetic Algorithm Based on Polygamy. In: *Proceedings of Third International Symposium on Intelligent Information Technology and Security Informatics*, pp. 371–373 (2010) ISBN: 978-0-7695-4020-7
5. Al Jaddan, O., Rajamani, L., Rao, C.R.: Improved Selection Operator for GA. *Journal of Theoretical and Applied Information Technology*, 269–277 (2005)
6. Sivanandam, S.N., Deepa, S.N.: *Introduction to Genetic Algorithms*. Springer, Heidelberg (2007) ISBN 9783540731894
7. Affenzeller, M., Winkler, S., Wagner, S.: *Genetic Algorithms and Genetic Programming: Modern Concepts and Practical Applications*. Chapman-Hall, CRC (2009) ISBN 1584886293
8. Rechenberg, I.: *Evolutionsstrategie – Optimierung technischer Systeme nach Prinzipien der biologischen Evolution* (PhD thesis). Fromman-Holzboog Verlag, Stuttgart (1973)
9. De Jong, K.A.: *An Analysis of the Behavior of a Class of Genetic Adaptive Systems* (Doctoral dissertation, University of Michigan) *Dissertation Abstracts International* 36(10), 5140B University Microfilms No. 76/9381 (1975)
10. Schwefel, H.P.: *Numerical Optimization of Computer Models*. John Wiley & Sons, New York (1981)
11. Goldberg, D.E., Deb, K.: A Comparative Analysis of Selection Schemes Used in Genetic Algorithms. In: *Foundations of Genetic Algorithms*, vol. I, pp. 69–93. Morgan Kaufmann (1991)
12. Fogel, D.B.: *Evolutionary Computation*. In: *Toward a New Philosophy of Machine Intelligence*. IEEE Press, Piscataway (1995)
13. Kumar, R., Jyotishree: Blending Roulette Wheel Selection & Rank Selection in Genetic Algorithms. In: *Proceedings of International Conference on Machine Learning and Computing*, vol. 4, pp. 197–202. IEEE Catalog Number CFP1127J-PRT (2011) ISBN 978-1-4244-9252-7
14. Baker, J.E.: Adaptive Selection Methods for Genetic Algorithms. In: *Proceedings of International Conference on Genetic Algorithms and Their Applications (ICGA1)*, pp. 101–111 (1985)

Gang Scheduling Strategy for Request Processing in Cluster Based Video-on-Demand Systems

A. Vinay¹, K. Bharath², and T.N. Anitha³

¹ Department of Information Science and Engineering, P E S Institute of Technology

² Department of MCA, B M S College of Engineering

³ Department of Computer Science and Engineering, S J C Institute of Technology
a.vinay@pes.edu

Abstract. The recent advances and development of inexpensive computers and high speed networking technology have enabled the Video on Demand (VoD) application to connect to shared-computing servers, replacing the traditional computing environments where each application was having its own dedicated computing hardware. The VoD application enables the viewer to select, from a list of video files, their favourite video file and watch it at their will. As VoD is becoming more ubiquitous, users are expecting more features. The overall objective of a VoD service provider is to provide a better Quality of Service (QoS). There are many ways to achieve higher level of QoS. One such solution is to avoid client starvation period / time by using effective scheduling strategy. This paper proposes a gang scheduling framework for video request processing in Video-on-Demand Systems. The simulation results show that the number of requests being rejected is 8.5%.

Keywords: Video-on-Demand (VoD) systems, Architectures and topology, Load balancing, Average rejection rate, Gang Scheduling.

1 Introduction

With the explosive growth of the Internet and increasing power of personal computers, interest has grown in a whole new class application called Video-on-Demand (VoD), where clients can request media at any time for immediate viewing. When a user wants to watch a video, he or she simply makes a selection from a list of available titles and within a few seconds, the video is ready to start. This simple interface requires many complicated network mechanisms that remain invisible to the user. The request routing mechanism redirects each user's request to the most appropriate server according to a variety of metrics such as distance, network load, or content availability. Depending on the technique used for content delivery, the server transmits the selected content to the user, either via a dedicated unicast stream or through a multicast connection, where many clients receive the same stream. The approach adopted for content allocation (where storage/streaming devices and video files are placed in the network) influences both the request routing and content delivery. The large size of video files makes it expensive and impractical to replicate the entire library at each site, and generally, only a fraction of the most popular titles

or file prefixes are stored at each replica. Distributing these files to the replicas consumes significant bandwidth. Therefore, the mechanism for content distribution, that is, the process of sending content from the origin server (library) to the replicas, must be designed carefully. Distribution is either triggered by a user's request or executed periodically to update the cache. At a higher level in the network hierarchy, mechanisms must exist for content ingestion that govern how and when new content is added to the network. The rate of content ingestion can vary significantly depending on the nature of the content provided by the VoD system. A library consisting predominantly of movie and documentary content might be expected to grow slowly, with a few titles added every day, whereas a library with significant TV content can grow at a very rapid rate. It is very likely that the usage and content of a VoD network will evolve over time, so a design must be robust to enable changes in usage patterns and ingestion rates.

However, the maximum number of concurrent streams that a VoD server can support is limited because of the constrained bandwidth. Moreover, when service demand surges at one time traffic exhibits a fairly higher value compared to average traffic. So, quality of service (QoS) assurance for video delivery is still a major concern. Scheduling video requests among servers is one of the most important issues that need to be addressed in QoS assurance. Scheduling involves the allocation of resources and times to tasks in such a way that certain performance requirements are met. Scheduling has been perhaps the most widely researched topic within real-time systems. This is due to the belief that the basic problem in real-time systems is to make sure that tasks meet their time constraints.

In this paper, a simulation model consisting of two homogeneous clusters is considered. The workload consists of video requests batch for same video (gangs) and high priority requests which can overtake gangs. Performance comparison of the proposed algorithm with a modified version which implements migrations under various workloads is made. Furthermore, reservation and aging techniques are used in order to regulate the number of migrations.

The rest of the paper is organized into various sections as follows: Different types of Video-on-Demand systems architectures are focused in section 2, related works in the area of Scheduling in Video-on-Demand system are presented in Section 3. Section 4 discusses the functional models and proposed gang scheduling method. Section 5 evaluates the Gang Scheduling technique in VoD system through extensive analysis and simulation. Section 6 concludes the work done.

2 VoD Architectures and Topology

For the real-time nature of streaming service, many system architectures have been proposed to provide scalability, QoS and fault-tolerance [4]. Current VoD architectures can be classified into four categories namely centralized, proxy-based, Content Distribution Network and hybrid. In a centralized architecture, the origin server is responsible for serving all the clients (Fig. 2a). Although this approach is simple to implement, it has serious weaknesses: a single point of failure and a high load on both the origin server and the surrounding network links and switches. These shortcomings have prompted the development of distributed architectures with proxy

servers installed at strategic locations in the network (closer to the clients). The proxy servers, located close to the user end, cache content to reduce the load on the origin, as shown in Fig. 2b. A request from a client is served by a proxy if it has a copy of the requested file; if it does not, the file is streamed from the origin. An extension of the proxy server approach is the use of content distribution networks (CDNs). In these networks, requests can be referred to other CDN servers (replicas or surrogates) that are generally located at the edge of the network core (Fig. 2c).

Hefeeda et al. argue that proxy-based approaches shift the bottleneck from the origin to the proxy servers and are not cost-effective solutions for streaming media [3]. Optimization exercises using reasonable, although necessarily approximate, network cost models contradict this argument, indicating that there is potential for substantial savings by deploying proxy servers at local network exchanges, provided there is sufficient demand [4]. A major reason for these design results is the rapid reduction in the cost of memory in recent years that has led to a shift in the cost trade-off between replica deployment and bandwidth consumption on network trunks. Hefeeda et al. propose a hybrid architecture based on the peer-to-peer (P2P) paradigm to distribute the files to the users (Fig. 2d), where the origin acts as a seed peer to help to route the requests and search for content [3]. Hybrid architectures are an extremely promising paradigm, because the storage and streaming resources scale (approximately) with the number of users. Thus, the architectures have the potential to adapt to growth in library size and usage. The details of such systems, including numerous operational issues and pricing considerations, remain to be resolved.

3 Related Work

Most of the existing systems follow first-come-first-serve (FCFS) scheduling technique where, as soon as the bandwidth for some server becomes available, the batch holding the oldest request with the longest waiting time is served immediately. An alternative is to use maximum-queue-length-first (MQL) [2] where in the batch with the largest number of pending requests (i.e., longest queue) is chosen to receive the service. FCFS offers fairness since the scheme treats each user equally regardless of the popularity of the requested video. This scheme, however, yields low system throughput because it may choose to serve a batch with fewer requests first while cause another batch with more requests to wait. To address this issue, MQL, which also maintains a separate waiting queue for each video, delivers the video with the longest queue (i.e., the largest number of pending requests) first. This policy maximizes server throughput, but is unfair to the users who request less popular videos.

Maximum-factored-queued-length first (MFQL) [4] attempts to provide reasonable fairness as well as high server throughput. This scheme also maintains a waiting queue for each video. When a server channel becomes free, MFQL selects the video vi with the longest queue weighted by a factor $1/fi$ to deliver, where fi denotes the access frequency or the popularity of the video vi . The factor fi prevents the server from favouring the popular videos at all times. However, it was observed that MFQL is not fair in most situations because it is solely determined by the queue length of the video. It is well known that popular movies always have the longer queue length than the others, and the effect of the factor fi is much smaller than the queue length.

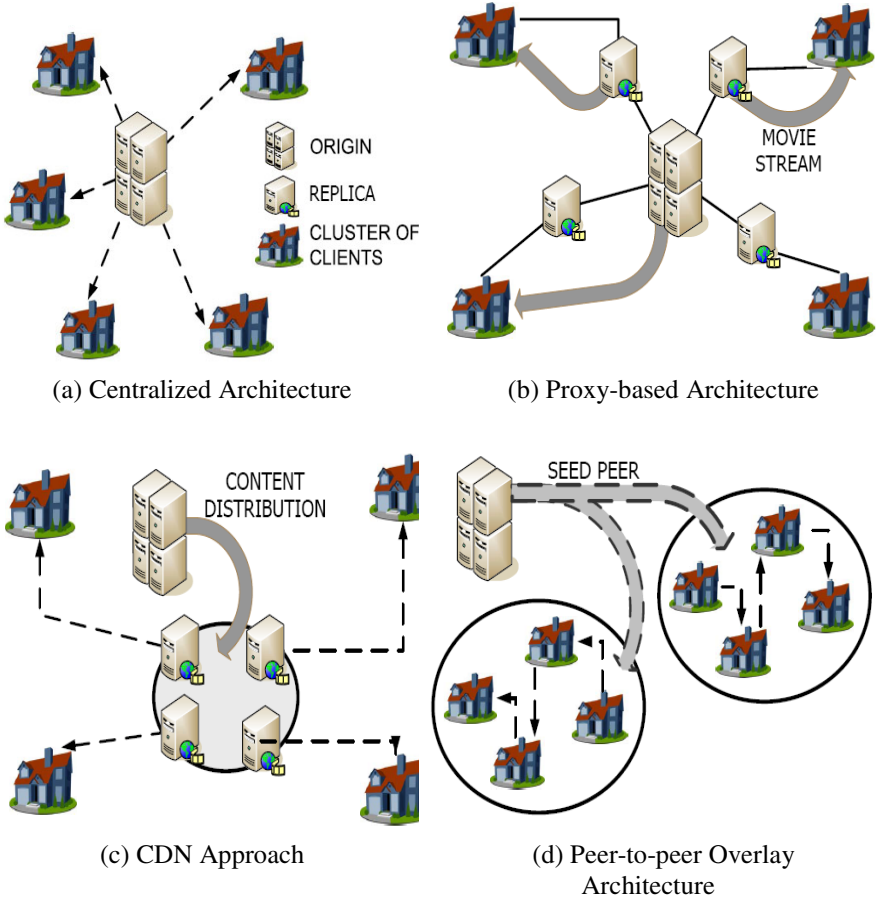


Fig. 2. Typical Video-on-Demand (VoD) System architectures

4 Proposed Method

VoD systems consisting of multiple server clusters have emerged as a solution to the increasing demand from millions of clients. The scheduling strategy that might be used in such a system is of great significance since the performance achieved is proportional to the algorithm's effectiveness. In a VoD system the scheduling algorithm is responsible for allocating servers to incoming requests. In a VoD environment most of the requests are those having high popularity i.e. the probability of clients requesting same video is more likely. Gang scheduling is an effective approach of scheduling such requests. Gang scheduling relies on time-space sharing. The main concept of this technique is to group the requests for same video, thus formulating a gang, and serve them simultaneously from different servers. Another major issue is the workload composition. There is a great impact on system performance when high priority requests exist in the workload. It might be necessary

for a high priority request to interrupt the current execution in order to satisfy its immediate requirements. In this way, the processing of a gang is delayed so as to maintain a certain Quality-of-Service (QoS) level in favor of a high priority request.

4.1 System Model

A simulation model is used to evaluate the system's performance. Simulation models enable us to study desired algorithms in detail. Workload traces would also be useful for the performance evaluation. However, such traces currently do not exist. Therefore, workload models are commonly used to evaluate an algorithm's performance. The simulation model consists of two server clusters. Each cluster is a homogeneous system consisting of P processors, each serving its own video request queue (Figure 3). Here it is assumed that the servers in each cluster are interconnected via a high speed local area network. The clusters are connected to each other via a wide area network. The communication between the processors is contention-free. Thus, we consider that the communication latencies are included implicitly in the gangs' execution time. However, overheads, which occur when migration schemes are applied, are considered. The service time of a video request is exponentially distributed with mean of $1/\mu$. The workload consists of batched video request (gangs) and high priority video requests. Gangs require a number of servers equal to its number of requests. The number of video requests are uniformly distributed in the range of $[1..P]$, where P is the number of processor in the system. Thus, the mean number of tasks per gang is equal to $(1+P)/2$. The mean inter-arrival time of gangs is exponentially distributed with a mean of $1/\lambda_1$.

Priorities can also be assigned to requests. Hence, high priority requests are also considered. These request need to start their execution immediately after their arrival. In order to achieve the immediate start-up of high priority request, an occupied server by a request previously arrived might have to be interrupted. If a currently streaming video request is interrupted, all the progress made is lost and streaming must be restarted. Furthermore, since a high priority request cannot be interrupted, that leaves us only with the choice of a request belonging to a gang. The mean inter-arrival time of high priority request is exponentially distributed with a mean of $1/\lambda_2$.

4.2 Request Routing

When a video request arrives, the request is dispatched to one of the available server clusters. This is accomplished by a *global dispatcher*. The global dispatcher that is used assigns request to the clusters in a uniformly distributed manner. The probability that the request is assigned to one of the clusters is equal i.e. 50% probability in a system consisting of two clusters. Here it is assumed that the global dispatcher does not have a priori knowledge about the clusters' current load. In this way, any overhead which could result from the implementation of a cluster information feedback algorithm is avoided. Therefore, a probabilistic algorithm is considered suitable for this purpose. After a parallel request has been send to a cluster, the *local dispatcher* assigns its requests to the available queues. The incoming requests are

distributed to the queues based on the shortest job queue policy. When a high priority request arrives at a cluster, it is routed based on the shortest job queue policy. The shortest queue of the cluster is selected for the request to join it, provided that there is no other high priority request occupying the aforementioned server.

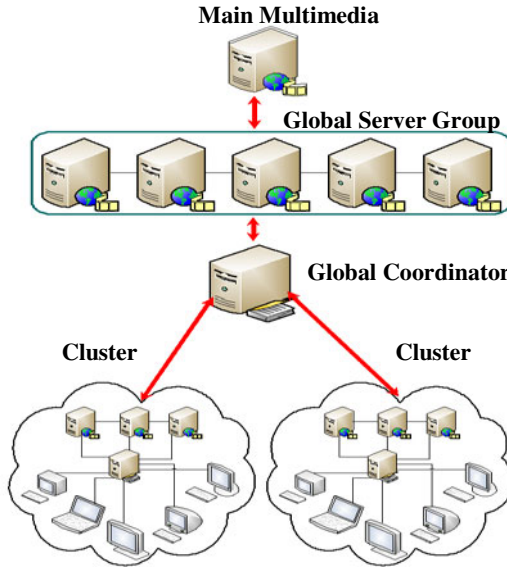


Fig. 3. Hybrid Video-on-Demand (VoD) System Model

4.3 Replication

In a VoD environment most of the requests are those having high popularity i.e. the probability of clients requesting same video is more likely. Gang scheduling is an effective approach of scheduling such requests. Gang scheduling relies on time-space sharing. The main concept of this technique is to group the requests for same video, thus formulating a gang, and serve them simultaneously from different servers. An enhanced version of First-Come-First-Served (FCFS) algorithm is applied to each cluster separately. According to this method, a request is scheduled whenever servers assigned to its tasks are available. When a request which is waiting in front of the queues cannot start streaming, the enhanced FCFS policy schedules other requests which are behind the aforementioned request. If the processing of a gang, which was scheduled using the modified FCFS method, is interrupted by a high priority request, then the current streaming is stopped and is rescheduled for execution at the head of their associated queues. Furthermore, any progress made for the specific gang is lost and must be redone.

4.4 Request Migration

A common problem that occurs when using gang scheduling is that there might be few servers which remain idle even though there are requests in their queues. This is due to the fact that the execution of a gang does not start unless all requests to a same video can start their execution simultaneously. Thus, the main reason for the delay of gang execution is that one or more requests are waiting in queues which belong to busy or reserved servers. Implementing migrations is a way to avoid starvations. Migration involves the transferring of a request from one remote queue to the head of a local queue. However, migration should be exercised with caution because, migration causes an overhead. Migration can be local or global. A local migration is the process of transferring request from one server queue to another queue belonging to the same cluster. A global migration is the process of transferring of one request from one cluster to another.

Scheduling algorithm might fail to schedule a gang due to the fact that there is no one-to-one mapping between its requests and servers. Whenever this kind of situation occurs, a need for local migration occurs. In local migration, gangs having one or more waiting request at the head of an available server's queue are examined. From these gangs one that requires the least number of migrations is selected. Any parallel requests exceeding the number of available servers are excluded from this procedure. The migrated requests are transferred to the head of the queues. Consequently, the video streaming can start immediately once the migration is completed. If any available servers still remain, we examine which gangs have a request waiting on the head of their queue. Provided that there are enough available servers in the other cluster, we select the gangs which require the smallest number of migrations. During a migration period, the target server is reserved in order to prevent other tasks from seizing it. By reserving the target server we ensure that after the migration has finished, the gang will have an available server for all of its requests and streaming will start immediately. The only possible cause that might hinder the streaming is the arrival of a high priority request.

When a high priority request arrives at a queue, the request that is occupying the corresponding server is interrupted in order to give its place to the high priority request. This means that the servers previously occupied by the gang are only allowed to serve high priority request. In this way, when there are no high priority requests occupying the necessary server, the reserved gang is brought back and served immediately without having to wait other gangs to finish their execution first. The reservations method saves the system from excessive network traffic due to migrations, does not let the requests to starve and keeps the response time of the migrated gangs low. Furthermore, the reservations technique allows the gangs which have migrated requests in a remote cluster to restart streaming video requested after being interrupted. We also make use of aging, in order to regulate the number of migrations which occur in the system. A queue is not available for other requests to migrate when there are requests in the queue which have already granted priority three times in the past. In this way, the starvation of the tasks belonging to this queue is prevented.

5 Results and Discussions

5.1 Performance Metrics

In order to evaluate the system's performance we employ the following metrics.

- *Response time* r_j of a parallel request j is the time interval from the arrival of the request to the global dispatcher to the service completion of this request. Note that since the global dispatcher does not use a queue to route the requests, we consider that there is no waiting on the global dispatcher.
- *Slowdown* s_j of a gang j is the response time of this request by the service time. This metric is used to measure the delay of a request against its actual service time. If e_j is the service time of the request j , then the slowdown is defined as follows:

$$s_j = r_j / e_j$$

Let m be the number of the total processed parallel request. The following metrics used for performance are defined as follows:

- The average response time RT :

$$RT = \frac{1}{m} \times \sum_{j=1}^m r_j$$

- The average slowdown SLD :

$$SLD = \frac{1}{m} \times \sum_{j=1}^m s_j$$

Additionally, the response time and the slowdown of each request are weighted with its size. In this way, it is avoided that request with the same service time, but with different number of parallel request, have the same impact on the overall performance. Let $p(x)$ represent the number of servers required by request x . The following weighted metrics are used:

- The average weighted response time WRT :

$$WRT = \frac{\sum_{j=1}^m p(x_j) \times r_j}{\sum_{j=1}^m p(x_j)}$$

- The average weighted slowdown $WSLD$:

$$WSLD = \frac{\sum_{j=1}^m p(x_j) \times s_j}{\sum_{j=1}^m p(x_j)}$$

5.2 Cost Function

We can express the total cost, C_{TOT} , as the sum of the cost of infrastructure, C_T , and the cost of transport, C_S .

$$C_{TOT} = C_T + C_S$$

The cost of infrastructure, C_T , includes the software and start-up cost of a location (A_i) and the cost of VoD servers (B_i) for every replica site i and the origin server. We express C_T as a function of the number of VoD servers installed at location i , n_i , and the origin, n_o .

$$\begin{aligned} C_T &= \sum_{i=0}^N A_i + B_i n_i \\ &= f_1(n_o) + \sum_{i=0}^N f_1(n_i) \end{aligned}$$

The cost of transport consists of two components: transport from the origin to replicas and clients, C_{SORi} , and transport from replica i to client i , C_{SRCi} . It includes the cost of node interfaces (C_{IF}) and of fiber (C_f). The transport from replicas to the user-end (small distances) uses direct fiber whereas the transport from the origin to the replicas uses DWDM connections.

$$\begin{aligned} C_S &= \sum_{i=1}^N C_{SORi} + C_{SRCi} \\ &= \sum_{i=1}^N f(n_{ORi}) + f(n_{CRi}) \end{aligned}$$

Where

$$C_{SRCi} = n_{RCi} \times (2 \times C_{IF} + d_{RCi} \times C_f)$$

$$\begin{aligned} C_{SORi} &= n_{ORi} \times (2 \times C_{IF}) + \frac{n_{ORi}}{dw_{max}} \times \\ &\quad \left[2C_{DWDM} + d_{ORi} \times C_f + \left(\frac{d_{ORi}}{\max_{amp}} \right) \times C_{LA} \right] \end{aligned}$$

The notations used in the above equations are summarized in the following table 1.

Table 1. Nomenclatures Used

n_{ORi}	:	Num. of interfaces (fibers) toward the origin.
n_{RCi}	:	Num. of interfaces (fibers) toward the user-end.
c	:	Fiber capacity. (Gbps)
C_{IF}	:	Node switch interface cost. (\$)
C_f	:	Cost of fiber. (\$/km)
C_{DWDM}	:	Cost of DWDM equipment (\$)
w_{max}	:	Number of fibers supported by DWDM equipment.
C_{LA}	:	Cost of line amplifier. (\$)
d_{amp}	:	Max. distance between two amplifiers. (km)

5.3 Simulation

The simulation of the above proposed gang scheduling technique was done using the concept of multithreading in Java. The simulation was carried out for 12000 seconds. Nearly 1650 requests were generated. The algorithm scheduled parallel requests to the servers of the cluster. Figure 5 illustrates the number of request being served from the systems. 96% of the total requests were served.

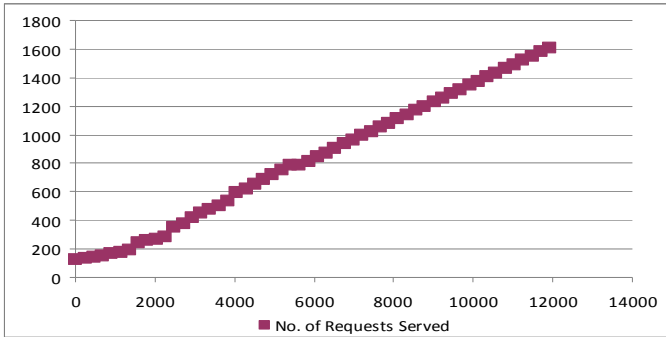


Fig. 5. Number of Requests Served

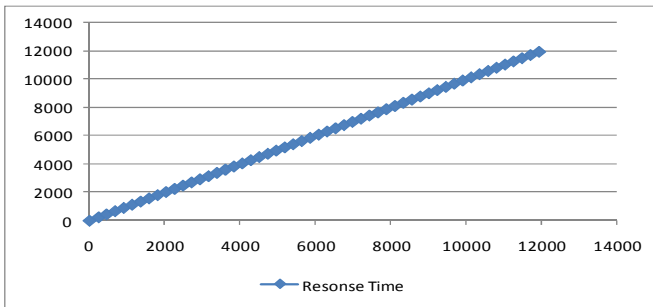


Fig. 6. System Response Time

Figure 6 illustrates the fact that the time duration between request arrival and start-up of response was very minimal .i.e. a good response time was exhibited. Similarly, an acceptable amount of service time was utilized by the system (Figure 7). Figure 8 illustrates the average slow down time which is ignorable in this case.

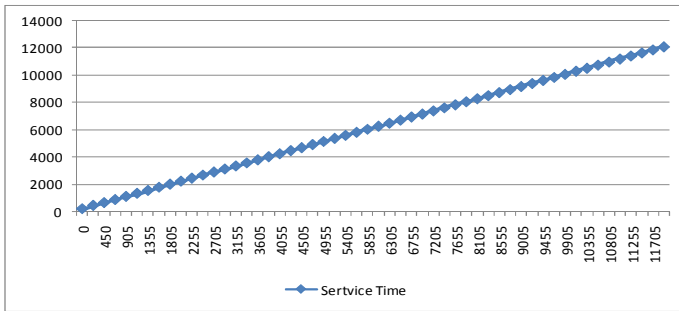


Fig. 7. System Service Time

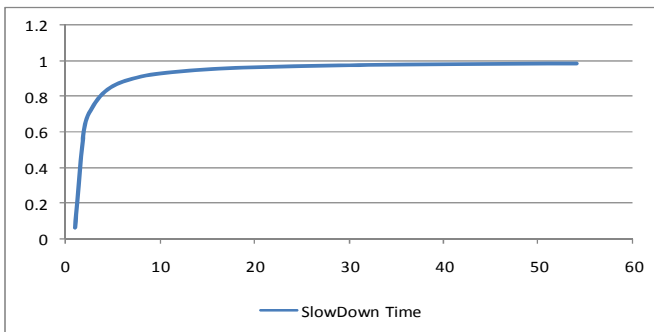


Fig. 8. Average Slowdown Time

6 Conclusions

To maximize the performance of VoD system and to avoid longer starvation of the clients, it is crucial to ascertain the amount of resources to be awarded to a particular client. This paper proposed an efficient gang scheduling technique promises a desirable QoS. The impact of migrations on the performance of gang scheduling strategy in a two-clustered VoD system with the presence of high priority jobs in the workload was examined. Experiments were conducted using a simulation model under various workloads. The results indicate a performance boost due to implementing the suggested gang scheduling algorithm with migration technique. Even in the case of increased high priority jobs workload implementing migrations achieved better performance.

References

1. Ghose, D., Kim, H.J.: Scheduling video streams in video-on-demand systems: A survey. In: The Proceedings of Multimedia Tools and Applications, ch. 11, pp. 167–195 (2000)
2. Notre-Dame de la Paix, Namur, “Video-on-Demand over Internet: A Survey of Existing Systems and Solutions” (2008)

3. Bar-Noy, A., Ladner, R.E.: Windows Scheduling Problems for Broadcast Systems. In: The Proceedings of SIAM Journal on Computing (SICOMP), pp. 1091–1113 (2003)
4. Ghandeharizadeh, S., Huang, L., Kamelm, I.: A Cost Driven Disk Scheduling Algorithm for Multimedia Object
5. Tsai, C.-H., Chu, E.T.-H., Huang, T.-Y.: WRR-SCAN: A Rate-Based Real-Time Disk-Scheduling Algorithm. In: The Proceedings of EMSOFT 2004, pp. 1–8 (2004)
6. Reddy, A.L.N., Wyllie, J., Wijayarathne, K.B.R.: Disk scheduling in a multimedia I/O system. The ACM Transactions on Multimedia Computing, Communications, and Applications - TOMCCAP 1(1), 37–59 (2005)

Does Social Communicability Mediate the Role of Trust in Mobile Phone Adoption? An Individual Level Multi-nation Exploratory Study

Kallol Bagchi and Somnath Mukhopadhyay

The University of Texas at El Paso, 500 West University Drive, Suite 205, COBA, 79968, USA
kbagchi@utep.edu, smukhopadhyay@utep.edu

Abstract. This paper investigates the role of Social Communicability Index (SCI) in the relationship between trust and Mobile Phone Adoption. Using individual-level secondary data from various nations from a reputable data base and controls such as age, gender and education level, we ran regressions and PLS and found that SCI of individuals does moderate the role of trust in mobile phone adoption for most nations. We conclude with future research direction.

Keywords: Mobile Adoption, SCI, Trust, Regression, Mediation, PLS.

1 Introduction

Trust is a belief that the sincerity or the good will of others can be generally relied upon and a function of degree of risk inherent in a situation [1] [2]. Past research mentioned that relations of trust can be viewed as resources for individuals [3]. Every commercial transaction has an element of trust [4]. Trust is, therefore, a fundamental concept and understanding of trust is essential in business situations.

Trust has been studied by many at both individual level [5] [6] in context of the Technology Adoption Model (TAM) and at a group/national level [7] [8]. At a group/national level, for example, previous research [9] found that trust and civic norms are stronger in nations with higher and more equal incomes. At an individual level, appropriate feedback mechanisms in electronic markets can induce credibility trust between two transaction parties [10]. There are several factors (including societal ones) that are responsible for how much people trust one other [11]. The factors are - individual culture and religion, how much an individual lived in a community, recent personal history of misfortune, the perception of being a part of a discriminated group and, several characteristics of the composition of one's community. The present study is at an individual level of many nations. The study involves a new social factor such as social communication index (SCI) of an individual in a given society that may play a role influencing trust and technology adoption relationship.

1.1 Trust and Information and Communication Technology (ICT) Adoption

Uncertainty and incomplete product information make trust critical in many situations [12]. Consumers may have to take a risk when they adopt a new product. Because

adoption involves high degree of risk and uncertainty, and the amount of information is limited, the preference for and use of information would be influenced by the individual's trusting thresholds. Trust is a prerequisite for individual social behavior including decision making [13]. So ICT adoption of individuals is expected to be governed by trust as it involves decision making. Individuals in high trusting societies tend to be more open-minded in searching for information and in the choice of new innovations. Because an individual's trusting threshold is the result of socialization into a society [14], trust and SCI could be of interest in ICT adoption.

Adoption of a technology can thus be regarded as a trusting process, with trust that is related to the quality and value of the technology product being adopted. The current study uses [15] construct to explore the influence of trust on adoption behavior in various nations.

The direct role of trust in technology adoption has been investigated in the past, both at an individual and national level [16] [17]. Thus research has generally found that individuals who have confidence in an ICT were more likely to adopt newer technologies. Interpersonal trust has a statistically significant influence on levels of Internet penetration [16]. They postulated that a 5% increase in trust ratings would result in a national growth rate in Internet subscribers of approximately 4% to 6.25%. Trust amongst stakeholders is critical to the success of outsourced information systems development projects [18]. In a virtual team, the propensity to trust had a robust impact [17]. Trust has proven to be an important variable influencing technology adoption. Recent research [19] found that communication's effect on individual performance is through trust. However, the role of social communicability in the relation between trust and ICT adoption was rarely investigated in ICT literature.

1.2 SCI

The diffusion of innovation theory explains the adoption process of an innovation as a process of information exchange, which is facilitated by mass media and interpersonal channels within the social system. One can use the theory to analyze how newly introduced technologies are communicated, evaluated, adopted or rejected, and re-evaluated by consumers [20]. The information flow through the communication channels (especially the word-of-mouth) is important in adopting an ICT. High uncertainty results in low trust [14]; which is associated with slowing down of the dissemination of information through channels of communication [21] [22]. For high uncertainty avoiding or low trust people, the use and flow of information among people is inhibited [23]. Since, the adoption of technology is highly dependent on relationship-based channels of communication through which information flows, a relationship between individuals with a high level of trust and ICT adoption intention should be influenced by communication flow. We investigate at an individual level, the role of SCI (additive, mediating or moderating) on the relationship between trust and ICT adoption for a set of nations. The central research question is as follows:

R1. What role (moderating, mediating etc.) does SCI play on the relationship between trust and ICT adoption among people within various nations?

2 The Theoretical Model and Hypothesis Development

2.1 The Additive Model

The above discussion leads to the conclusion that both trust and SCI may play direct roles in ICT adoption such as mobile telephony.

H1. SCI and Trust effects are both additive on mobile phone adoption (MA)

2.2 The Moderating Model

According to [24], "In general terms, a moderator is a qualitative (e.g., sex, race, class) or quantitative (e.g., level of reward) variable that affects the direction and/or strength of the relation between an independent or predictor variable and a dependent or criterion variable." SCI can affect the relationship between trust and MA. The higher the trust and higher the SCI of an individual, the greater could be trusting and communicating ability of that individual in the society which could spread the good news about mobile phones to other people of the society which would result in higher MA.

H2. SCI moderates the relation between Trust and MA adoption or Trust exerts an indirect effect on MA through SCI.

2.3 The Mediating Model

Moderator variable is one that influences the strength of a relationship between two other variables, and a mediator variable is one that explains the relationship between the two other variables [24]. It is well-known that information about a product spreads more effectively through word-of-mouth networks than through external media such as TV or newspapers. If SCI is absent, word-of-mouth networks could be severely weakened resulting in poor MA.

H3. SCI mediates the relation between Trust and MA adoption

3 Data and Method

3.1 Data

We collected data from the European Social Survey (www.europeansocialsurvey.org or ESS in brief) conducted in 2007-2008. Data from 28 European nations were used with each nation containing more than 1000 data points (an overall total of more than 50,000 data points). Our technology of choice is mobile telephone.

The trust variable was defined by the generalized trust measured by the following three questions: trust in people, trust in human fairness and trust in human nature. These are- i) Most people can be trusted or you can't be too careful: 0=You cannot be too careful to 10=Most people can be trusted, ii) Most people try to take advantage of you, or try to be fair: 0= Most people try to take advantage to 10=Most people try to be fair, iii) Most of the time people helpful or mostly looking out for themselves: 0=People mostly looking out for themselves to 10= People mostly try to be helpful.

The correlations between question (i) and (ii) ($r=0.62$, $N=50,540$), (i) and (iii) ($r=0.54$, $N=50,776$) and (ii) and (iii) ($r=0.57$, $N=50,449$) were strong for the overall set thus indicating that these three questions can measure different related aspects of generalized trust (henceforth trust).

The SCI index was measured by the average of two variables defined in ESS based on the following questions: i) How often socially meet with friends, relatives or colleagues: 1=Less than a month to 5=Several Times a Week, ii) Take part in social activities compared to others of same age: 1=Much less than most to 5= Much more than most. SCI had scores ranging from 1 to 5 with 10 possible values.

MA data was also measured from the same database which was generated from the question item as: Personally (I) have (a) mobile telephone 1=yes, 2=no. Variables age, gender, and, educational level were also used from the same database as control variables.

3.2 Method

The additive model ran both trust and SCI as independent variables. For the additive model to hold, the requirements were to test for statistical significance of both independent variables. We multiplied trust with SCI and normalized the variable thus obtained for moderation effect. We then included this variable in a regression that contained both trust and SCI as independent variables. The moderation test examined the statistical significance of the product variables in the aforementioned regression model.

To examine mediation effects, we followed [24]. They have discussed four steps in establishing a complete mediation. Let X be the initial variable (that affects another variable), Y is the outcome variable and M is the mediator variable. Their suggested steps are – (1) Show that the initial variable is correlated with the outcome, (2) Show that the initial variable is correlated with the mediator, (3) Show that the mediator affects the outcome variable. Use Y as the criterion variable in a regression equation and X and M as predictors. Thus, the initial variable must be controlled in establishing the effect of the mediator on the outcome, (4) To establish that M completely mediates the X-Y relationship, the effect of X on Y controlling for M should be zero. The effects in both Steps 3 and 4 are estimated in the same equation.

Past study suggests that steps 1-3 (and not step 4) need to be satisfied only if there is no need to show complete mediation [25]. In the opinion of a majority of analysts, Step 1 is also not required [26]. Thus steps 2 and 3 have to be followed at a minimum to show the existence of some form of mediation (not the complete one). In our case, M is SCI, Y is the mobile adoption (MA) and X is the trust variable. Logistic regression methods were followed to test steps 2 and 3. Trust \rightarrow SCI regression was performed in OLS.

4 Results and Discussion

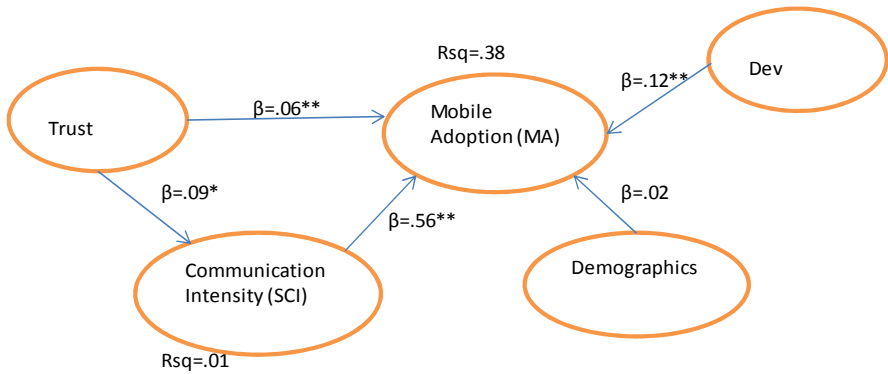
Table 1 shows the results of testing the models with control variables. The role of trust mediation becomes prominent over other possibilities. Thus R1 and H3 are answered in the affirmative whereas support for H1 and H2 are lacking

Next we ran the overall European region model (consisting of most European nations, 28 in total) in PLS controlling for demographics (Demo, which consisted of age, education level and gender) and emerging/developed nations (Dev=1 for developed nations, 0 for emerging nations). The PLS model for mediating impact was generated using Warp PLS 2.0 software (<http://www.scriptwarp.com/warppls>). The mediating model is based on the works of [27] [28]. We abbreviate as follows: latent variable *mobile* denotes MA; *demo* denotes latent variable formed from three demographic variables, age, gender and levels of education. The Rsq for MA was 0.39. All model fit indices had satisfactory values. Average path coefficients (APC) =0.172, $p < 0.001$ and Average Rsq (ARS)=0.195, $p < 0.001$ (the p-values were both $< .05$ as required) and Average VIF (AVIF)=1.168, was < 5 as recommended in [29].

Table 1. The Results from the Models (No. of Nations = 28)

Models	Significance of the Model with controls	Range of Nagelkerke R ²	Hypothesis supported
H1. Additive			No
Trust -> MA	2 out of 28		
SCI ->MA	18 out of 28		
No. of nations for which H1 holds	0 out of 28	.096-.493	
H2. Moderation			No
Trust -> MA	0 out of 28		
SCI ->MA	5 out of 28		
Trust*SCI -> MA	1 out of 28		
No. of nations for which H2 holds	1	.097-.493	
H3. Mediating			Yes
i)Trust -> SCI	26 out of 28		
ii)SCI-> MA	18 out of 28	.091-.493	
iii) Trust+SCI -> MA (Trust is controlled) for	18 out of 28 (coeff. Of SCI)		
No. of Nations for which H1, H2 did not hold but H3 did	18		

The measurement model construct validity (convergent and discriminant) [30][31][32] is next discussed: two of the Trust items loaded well with the exception of “trust in human nature” component, which loaded poorly. It was subsequently dropped. Similarly, age loaded well on Demo, but not the other two items, gender and educational level. However, these did not cross-load on other variables. The two components of SCI loaded well. The composite reliability (CR) of Trust was 0.78; however, the CR of Demo was not adequate (0.35). The discriminant validity was good with the square root of the average variance extracted (AVE) to be much larger than any correlation among any pair of latent constructs.



Legend: *: $p < .02$, **: $p < .01$

Fig. 1. The Mediating Model Results using PLS

The structural model (shown in Figure 1) supported the mediating hypothesis. Except for Demo, all coefficients were significant ($p < .01$). The Sobel’s standard error for mediating effect was .025. The T value for mediating effect was 2.06, significant at $p < .034$. Thus the overall European regional PLS model also showed that mediating effect of SCI was significant. The PLS mediating model was subsequently tried on two other nations separately, Sweden and the UK. The results obtained were very similar (with mediating effects as $T = 2.68$, $p < .008$ for Sweden and $T = 1.92$, $p < .06$, for the UK data set (with Sobel’s standard errors)). The moderating PLS model for these two nations also was not relevant (consistent with the results shown above in Table 1).

5 Conclusion

In summary, we found that (1) SCI and Trust effects are not additive on Mobile phone adoption, (2) SCI does not moderate the relation between Trust, and, Mobile phone adoption, and, (3) SCI does mediate (in most cases) the relation between Trust and Mobile phone adoption. The study contributes in several areas. First, a study investigating the role of SCI as a mediating variable between trust and ICT adoption is absent. Second a study involving multiple nations from a geographical region such

as the Europe is richer than a single-nation study. Third the results showed that SCI mediates the relationship between trust and ICT adoption in most nations in mobile adoptions. An interesting question is what happens with other types of technology adoption across nations and what role SCI plays in the relationship between Trust and ICT use. Future studies may look into that as well as testing the model for other world regions. The roles of moderating-mediation and mediating-moderation can also be investigated in future.

This preliminary study has a few limitations. First, the study is cross-sectional in nature and in a given year, several nations could be in different stages of adoption. Thus the role of stage is an important one. However, this can only be investigated with a mixed-level model structure. Second, there are several other variables that can contribute to technology adoption and Trust and SCI are two representative examples. Third the model may be improved by reconsidering the Demo construct with income variable and SCI construct with more/less variables. For example, in a preliminary additional test, SCI with a single item yielded much better results. Convergent validity for the mediating model may be improved with these changes. Binary nature of the dependent variable renders mediation model in SEM as exploratory, although based on same principle as the ordinary dependent variable [28] and so results need to be interpreted carefully.

References

1. Rotter, J.B.: A new scale for the measurement of interpersonal trust. *J. Pers* 35(4), 651–665 (1967)
2. Koller, M.: Risk as Determinant of Trust, *Basic and Applied Psychology* 9(4), 265–276 (1988)
3. Coleman, J.: *Foundations of Social Theory*. Harvard University Press, Cambridge (1990)
4. Arrow, K.: Gifts and Exchanges. *Philosophy and Public Affairs* 1(4), 343–362 (1972)
5. McKnight, D.H., Chervany, N.L.: What Trust Means in E-Commerce Customer Relationships: An Interdisciplinary Conceptual Typology. *International Journal of Electronic Commerce* 62, 35–59 (2002)
6. Gefen, D.: E-Commerce: The Role of Familiarity and Trust. *The International, Omega* (2000)
7. Zack, P., Knack, S.: Trust and Growth. *Economic Journal* (April 2001)
8. Fukuyama, F.: *Trust: The Social Virtues and the Creation of Prosperity*. The Free Press, New York (1995)
9. Knack, S., Keefer, P.: Does social capital have an economic payoff? *Quarterly Journal of Economics* 112, 1251–1273 (1997)
10. Ba, S., Pavlou, A.P.: Evidence of the Effect of Trust Building Technology in Electronic markets: Price Premiums and Buyer Behavior. *MIS Quarterly* 26(3), 243–268 (2002)
11. Alesina, A., La Ferrara, E.: Who trusts others? *Journal of Public Economics* 85(2), 207–234 (2002)
12. Swan, J.E., Nolan, J.J.: Gaining customer trust: a conceptual guide for the salesperson. *Journal of Personal Selling & Sales Management*, 39–48 (November 1985)
13. Luhmann, N.: *Trust and Power*. Wiley, Chichester (1979)
14. Hofstede, G.: *Culture's Consequences*, 2nd edn. Sage Publications, Thousand Oaks (2001)

15. Inglehart, R., Baker, W.: Modernization, Cultural Change and the Persistence of Traditional Values. *American Sociological Review* 65, 19–51 (2000)
16. Huang, K.C., Leland, J., Shachat, J.: Trust, the Internet, and the digital divide. *IBM Systems Journal* 42(3), 507–518 (2003)
17. Jarvenpaa, J.S., Leidner, D.E.: Communication and trust in global virtual teams. *Organization Science* 10(6), 791–815 (1999)
18. Lander, M.C., Purvis, R.L., McCray, G.E., Leigh, W.: Trust-building mechanisms utilized in outsourced IS development projects: a case study. *Information & Management* 41, 509–528 (2004)
19. Sarker, S., Ahuja, M., Sarker, S., Kirkeby, S.: The Role of Communication and Trust in Global Virtual Teams: A Social Network Perspective. *Journal of Management Information System* 28(1), 273–309 (2011)
20. Rogers, E.: *Diffusion of Innovations*, 4th edn. The Free press, New York (1995)
21. Anderson, E., Weitz, B.: Determinants of Continuity in Conventional Industrial Channel Dyads. *Marketing Science* 8(4), 310–323 (1989)
22. Nakata, C., Sivakumar, K.: National Culture and New Product Development: An Integrative Review. *Journal of Marketing* 60(1), 155–165 (1996)
23. Deshpande, R., Zaltman, G.: A Comparison of Factors affecting Researchers and Manager Perceptions of Market Research Use. *Journal of Marketing Research* 24, 114–118 (1987)
24. Baron, R.M., Kenny, D.A.: The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology* 51, 1173–1182 (1986)
25. Kenny, D.A., Kashy, D.A., Bolger, N.: Data analysis in social psychology. In: Gilbert, D., Fiske, S., Lindzey, G. (eds.) *The Handbook of Social Psychology*, 4th edn., vol. 1, pp. 233–265. McGraw-Hill, Boston (1998)
26. Kenney, D.: On Mediation, <http://davidakenny.net/cm/mediate.htm> (accessed on September 24, 2011)
27. Hayes, A.: Beyond Baron and Kenny: Statistical Mediation Analysis in the New Millennium. *Communication Monographs* 76(4), 408–420 (2009)
28. Hayes, A., Preacher, K.J.: Quantifying and Testing Indirect Effects in Simple Mediation Models when the Constituent Paths are Nonlinear. *Multivariate Behavioral Research* 45(4), 627–660 (2010)
29. Kock, N.: *WarpPLS 2.0 User Manual*. ScriptWarp Systems, Laredo Texas (2011), <http://www.scriptwarp.com/warppls/UserManual.pdf> (retrieved online September 24, 2011)
30. Straub, D., Boudreau, M.-C., Gefen, D.: Validation Guidelines for IS Positivist Research. *Communications of the Association for Information Systems* 13, 380–427 (2004)
31. Fornell, C., Larcker, D.: Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. *Journal of Marketing Research* 18(1), 39–50 (1981)
32. Hair, J., Anderson, R., Tatham, R.: *Multivariate Data Analysis*. Macmillan, New York (1987)

Data Security for Virtual Data Centers by Commutative Key

Maram Balajee¹, Challa Narasimham², and Y. Ramesh Kumar³

¹ Dept. of IT, G M R Institute of Technology,
Rajam, Andhra Pradesh, 532127, India
balajee.m@gmrit.org

² Dept of IT, V R Siddhartha Engg. College, Vijayawada
narasimham_c@yahoo.com

³ Dept. of IT, Avanathi Inst. of Technology,
Vizianagaram, AP, India
javaramesh143@gmail.com

Abstract. The maintenance cost of a Data Center (In Small Organizations) is very high and difficult also. So an economically better choice is to use cloud computing and cloud storage instead of manage data centers by itself. But in Cloud Storage, Cloud user's sensitive data is in the control of a third party. Here the customer can't trust the Cloud Storage. But Cloud storage providers' claims that they can protect the data, but no one believe them. So this paper presents a framework to ensure data security in cloud storage system. In this framework, we use Commutative property, Bitwise XOR for managing keys between cloud storage and customer. And UNICODE and Colors for encrypt and decrypt the data both in cloud storage and customer's system.

Keywords: UNICODE, Colors, commutative key, cloud storage, data center, bitwise XOR.

1 Introduction

As of now, many technologies have introduced to store the data in cloud storage. But in some industries, the data are being dynamically created. And the data sources are geographically distributed all over the world. But cloud storage has the potential of providing geographically distributed storage services since cloud can integrate servers and clusters that are distributed all over the world and offered by different service providers into one virtualized environment. This can potentially resist disastrous failures and achieve low access latency and greatly reduced network traffic by bringing data close to where they are needed.

Cloud computing can be defined as a type of parallel and distributed system consisting of a collection of interconnected and virtualized computers that are dynamically provisioned, and presented as one or more unified computing resources based on service-level agreements established through negotiation between the service provider and consumers [1].

cloud storage

Cloud storage is a new distribution model, however, with the potential for economies of scale. Aside from cost, its benefits are outsourced operation, simple, unlimited growth and 'enterprise' features for smaller users - like high availability, security, data protection, etc.

There are nearly as many definitions of cloud storage as there are providers of cloud services. In simplest terms, cloud storage is data storage or services hosted remotely on servers and storage devices on the Internet or a similar private network, usually hosted by a third party.

Cloud storage is a subset of cloud computing, in which the term cloud refers to the wide area network infrastructure, including switches and routers, for a packet-switched network. When capitalized, cloud usually refers to the public data network, including the Internet.

Cloud computing has probably been best defined by the National Institute of Standards and Technology (NIST) as:

Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

Data is deployed to cloud storage either through Web-based applications or through Web services application programming interfaces (APIs). Web-based applications are often used for manual access to data or management functions, while APIs are used for more automated or transparent approaches. Since standard APIs and communications protocols are used, the physical location of the data becomes irrelevant, since it can be made available virtually anywhere via the Internet or private network. This also means that cloud data can be easily replicated to multiple locations for fault tolerance, high availability and other purposes, often without involvement of the customer.

2 Existing Systems

In cloud storage, the data is stored in Remote system which is owned by third party vendor. Whenever the customer wants to get data, the data will be transferred from third party vendor to customer. In this scenario, how the owner of the data trust the third party vendor. Because the plain data is available in the server which is being maintained by third party vendor. So the owner of the data can't trust the third party vendor. So there is a need of alternative solution. The following section explains that alternative solution.

2.1 UNICODE

ASCII which stands for American Standard Code for Information Interchange became the first widespread encoding scheme. However, it is limited to only 128 character definitions. Which is fine for the most common English characters, numbers and punctuation but is a bit limiting for the rest of the world? The people in the world naturally wanted to be able to encode their characters too.

So there is a need of a new character encoding scheme was needed and the UNICODE [2] standard was created. The objective of UNICODE [2] is to unify all the different encoding schemes so that the confusion between computers can be limited as much as possible. These days the UNICODE [2] standard defines values for over 105,000 characters and can be seen at the UNICODE [2] Consortium. It has several character encoding forms, UTF standing for UNICODE [2] Transformation Unit:

- UTF-8: only uses one byte (8 bits) to encode English characters.
- UTF-16: uses two bytes (16 bits) to encode the most commonly used characters.
- UTF-32: uses four bytes (32 bits) to encode the characters. UTF-32 is capable of representing every UNICODE [2] character as one number.

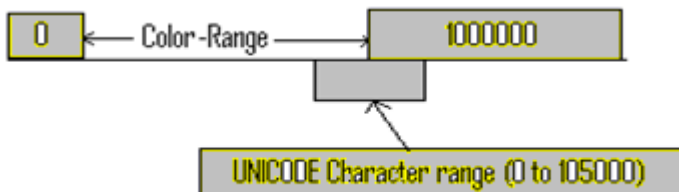
2.2 Colors Supported by Computer

Careful measurements of our visual system's best performance have been made by psychophysicists (people who study human responses, like seeing Color, to things in the world, like light). They have shown that we can see about 1000 levels of light-dark, 100 levels of red-green, and 100 levels of yellow-blue for a single viewing condition in a laboratory. This means that the total number of Colors we can see is about $1000 \times 100 \times 100 = 10,000,000$ (10 million). A computer displays about 16.8 million Colors to create full Color pictures, really more than necessary for most situations.

2.3 Cryptography with UNICODE and Colors

This method is fully based on Private-key cryptography. In secret-key cryptography schemes, a single key is used to encrypt & decrypt data. A secret key may be held by one person or exchanged between the sender and the receiver of a message. If secret-key cryptography is used to send secret messages between two parties, both the sender and receiver must have a copy of the secret key. The method is to send it via another secure channel or even via overnight express, but this may be risky in some cases.

In this method, the sender will decide the range of Colors which will be assigned to 1, 05,000 UNICODE [2] characters. The selection of Colors is in the following way:



In this proposed new policy, the sample binding between character / symbol/digit, UNICODE [2] and Color is in the following way:

Ch	UNICODE	COLOR
a	U+0061	Grey
b	U+0062	Red
c	U+0063	Green

Fig. 1. Sample mapping of Char/symbol/digit, UNICODE and Color

2.4 Use of Receiver-Key

A computer displays about 16.8 million Colors to create full Color pictures, really more than necessary for most situations. Now we are considering 10 million Colors only. And the UNICODE [2] standard defines values for over 105,000 characters and can be seen at the UNICODE [2] Consortium. Now 10 million Colors and 105,000 characters are available in a computer system.

Now we can create a dynamic mapping table like Fig 1. If we select starting position is 825,001. Then 825,001st Color is assigned to the first UNICODE [2] character. The 825,002nd Color is assigned to the second UNICODE [2] character. And so on. Finally the 930,000th Color is assigned to 105,000 UNICODE [2] character.

2.5 Encryption

This is kind of cryptography is fully based on UNICODE [2] and Colors. First of all, it checks each and every character in the given file. Then it finds concerned UNICODE [2] of each character is. Then it gives the corresponding Color for each UNICODE [2] character according to the predefined mapping.

According to the starting-point, the dynamic mapping table is created before encryption. Next, it takes the first character from text and finds the UNICODE [2]. According to the UNICODE [2], it will take concerned Color. Then it will take the 2nd character and so on. Now all characters are translated into corresponding Colors. It means, all characters are encrypted into Color-charts.

Now it takes the first character from shared receiver-key and finds the concerned Color. And overlap all Color-charts with this new Color. Then it takes the 2nd character, finding the concerned Color and overlaps recent Color-charts with Color and so on.

After encryption, the data is transferred from owner of the data to Cloud Storage. So the Cloud Storage is having encrypted data, which is not decrypted by the owner of the Cloud Storage also.

2.6 Decryption

Now the receiver receives encrypted data and temporary-key. After receiving a temporary-key from sender, the receiver calculates the sender-key. According to the

sender-key, the receiver prepares a mapping between alphabet/special character/digit, UNICODE [2] and Color. By using this chart, the receiver easily identify the COLOR->UNICODE->character.

3 Proposed System

The following Steps explain how Proposed Systems works.

Step 1: In this proposed method, there is a need to apply Commutative Property. According to Commutative Property, there is a need of one secret-key i.e Receiver-key and one public key i.e Intermediate-key. Here the sender creates Session-keys dynamically. Here no need to send any key with the message to the receiver.

Step 2: When the owner of the data wants to use Cloud Storage, then he/she simply encrypt the data by using the existing method (which is explained in existing systems) and shared secret-key i.e receiver-key. Now the Cloud Storage is having encrypted data only.

Step 3: When customer (Receiver) wants to get data from Cloud Storage, he/she has to send request. After receiving a request, a Session-key will be created and a Intermediate-key will be calculated based on session-key, shared secret-key i.e Receiver-key and bitwise XOR. And the encrypted data will be encrypted once again with Cloud's session-key.

Step 4: Now the data is transferred from Cloud Storage to Customer. Here no need to any key, because the Intermediate-key is public key. Now the Customer is having encrypted data, Intermediate-key and Receiver key.

Step 5: Based on available keys (Intermediate- key & Receiver-key), the receiver can calculate the required key i.e Cloud's Session-key. Here Intermediate-key is public-key.

Step 6: Now the receiver can decrypt the encrypted data by using Cloud's Sender-key and Receiver's receiver-key.

So no need to send any key while transmission of data from Cloud's Storage Centre and Receiver. Because the Receiver is having receiver-key and Intermediate-key is public key, so the receiver can calculate Cloud's Session-key by using Intermediate-key, receiver-key and bitwise XOR.

After calculating Cloud's Session-key, the receiver can decrypt the data by using Cloud's session-key and Receiver's receiver-key. So commutative property plays vital role in this proposed system.

3.1 The Importance of Commutative Property

According to Mathematics, the commutative property is $a.b=b.a$. According to Proposed system, a and b are sender-key, receiver-key respectively. By default, the data is encrypted with receiver-key (which is shared between End-user and owner of the data) i.e a. As and when required, the encrypted data is again encrypted with sender-key i.e. b.

In this way, sender is having sender-key and receiver is having receiver-key only. But the sender sends encrypted data & temporary key (t) to the receiver. Here t is bitwise XOR of a and b. So Temporary-key (t): $a \wedge b$.

Then the receiver receives encrypted data and temporary-key (t) only. Based on receiver-key and temporary-key (t), the receiver calculates sender-key (a) using bitwise XOR (\wedge).

While data transmission, the hacker can't get neither original data nor original keys (sender-key & receiver-key). In this way, commutative key provides more security to the data which is stored in Cloud Storage/ Data Center. So third-party vendor also not able to get the data, which is stored in Cloud Storage.

In the proposed method, we can take data from different languages in the world like English, French, German, Latin, Russian, Hindi, Telugu, Tamil, Kannada, Bengali, Malayalam, Urdu etc. And we can take sender key, receiver key from those languages also. So those keys would not be guessed by hackers.

4 Explanations with an Example

Suppose we want to encrypt the message "rajam" with shared secret key "abc". Now the above message is translated into the following UNICODE [2] characters:

rajam->U+0072 U+0061 U+006A U+0061 U+006D

4.1 Encryption

Then these UNICODE [2] characters are translated into the following Color-chart.



Fig. 2. Basic Color Chart of given Message

Here shared receiver-key is "abc". Assigned Colors are as follows:

Ch	UNICODE	COLOR
a	U+0061	grey
b	U+0062	red
c	U+0063	olive

In first iteration, the Basic Color Chart (BCC) is overlapped with the corresponding Color of the first character in shared receiver-key 'a'. It is called First Color Chart (First CC). In 2nd iteration, the recent Color chart overlapped with the corresponding Color of 'b' and so on. After completion of overlapping, the Final Color Chart (FCC) is looking like the following:



Then a temporary key is calculated based on sender-key, receiver-key and bitwise XOR in the following way. Assume,

Receiver-key: “abc”; Sender-key: “xyz”

Temporary key will be calculated by using Receiver-key, Sender-key and bitwise XOR in the following way:

```

01100001(a)  01100010(b)   01100011(c)
01111000(x)  01111001(y)   01111010(z)
.....
00011001(↓)  00011011(←)   00011001(↓)
.....
    
```

Here the result is bitwise XOR of “abc” and “xyz”. Now the temporary-key is “↓←↓”. This temporary-key also converted into Colors [3]. Now encrypted data and temporary-key are ready.

When receiver wants to get data, then the receiver will receive encrypted data and temporary-key (“↓←↓”). Now the receiver will calculate sender-key by using receiver-key, temporary-key and bitwise XOR. After performing bitwise XOR, the receiver will get sender-key.

By using sender-key and receiver-key, the receiver simply decrypts the received encrypted data. Initially the receiver should decrypt using receiver-key then sender-key in the following way:

After receiving the Color-chart, the receiver prepares Mapping between alphabet / special character / digit, UNICODE [2] and Color. According to receiver-key it is very simple to find equaling UNICODE [2] then character.

4.2 Decryption





After receiving encrypted data and temporary-key, the receiver calculates the sender-key by using receiver-key, temporary-key and bitwise XOR in the following way:

Receiver-key (“abc”) & temporary-key (“↓←↓”):

```

01100001(a)  01100010(b)   01100011(c)
00011001(↓)  00011011(←)   00011001(↓)
.....
01111000(x)  01111001(y)   01111010(z)
.....
    
```

Here the result is bitwise XOR (^) of receiver-key and temporary-key i.e sender-key (“xyz”). By using the mapping table like Fig 1, we can decrypt the encrypted message like the following:

In this the first cell  is converted into  by applying receiver-key. After applying sender-key, the cell  is converted into , which indicates ‘b’. If we apply same procedure to remaining cells then we can get the actual message “rajam”

5 Pictorial Representations

Here the owner of the data is encrypt the data by using shared secret-key i.e receiver-key. And it will be uploaded to Cloud Storage by using above said method i.e. UNICODE [2] AND colors [3] combination. And the owner of the data will calculate temporary-key based on sender-key, receiver-key and bitwise-XOR. Here sender-key and receiver-key are 2 components in Commutative property. The procedure is in the Fig3.

After receiving encrypted data and temporary-key, the receiver calculates sender-key using receiver-key, temporary-key and bitwise-XOR (^). Then receiver decrypt the data using both sender-key and receiver-key like the Fig4.

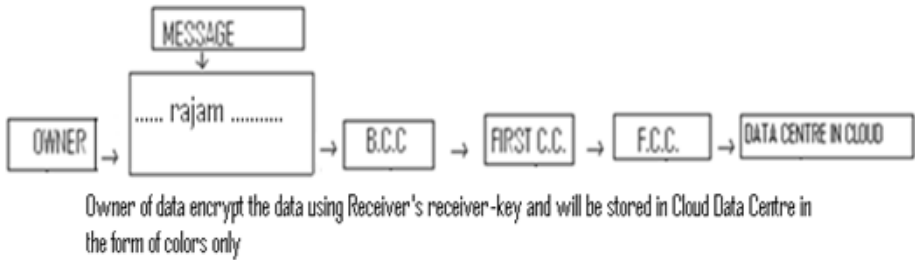


Fig. 3. Data Encryption and storage of data in Cloud's Virtual Data Center

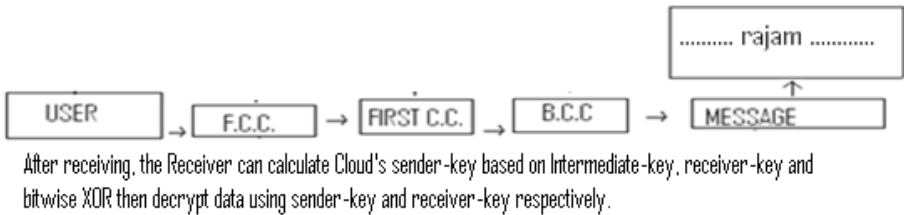


Fig. 4. Data Decryption using sender-key, receiver-key, bitwise XOR, UNICODE & Colors

References

- [1] Buyya, R., Yeo, C.S., Venugopal, S., Broberg, J., Brandic, I.: Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing as the 5th Utility. *Future Generation Computer Systems* 25(6), 599–616 (2009)
- [2] Balajee, M.: UNICODE [2] and Colors Integration for Encryption and Decryption. *IJCSE* 3(3) (March 2011)
- [3] Balajee, M., Narasimham, C.: IPVDD: Intrusion prevention for virtual Data Centers (A Framework for Encryption and Decryption). *IJCST* 2(4) (October- December 2011); ISSN: 0976-8491 (online), 2229-4333 (print)

Genetic Algorithm for Optimizing Functional Link Artificial Neural Network Based Software Cost Estimation

Tirimula Rao Benala¹, Satchidananda Dehuri²,
Suresh Chandra Satapathy¹, and S. Madhurakshara¹

¹ Anil Neerukonda Institute of Technology and Sciences
Sangivalasa, Visakhapatnam, Andhra Pradesh, India
b.tirimula@gmail.com, sureshsatapathy@ieee.org,
madhura_akshara@yahoo.co.in

² Department of Information & Communication Technology
Fakir Mohan University, Vyasa Vihar, Balasore-756019, India
satchi.lapa@gmail.com

Abstract. As Software becomes more complex and its scope dynamically increases, the importance of research on developing methods for estimating software development efforts has perpetually increased. Such accurate estimation has a prominent impact on the success of projects. The proposed work uses Functional Link neural network (FLANN) based estimation, which is essentially a machine learning approach, is one of the most popular techniques. In this paper the author has proposed a 2 step process for software effort prediction. In first phase known as training phase the FLANN selects the matching class (datasets) for the given input, which is improved by optimizing the parameters of each individual dataset by Genetic algorithm. In second step known as testing phase, the prediction process is done by Functional Link Artificial Neural Networks. The proposed method uses COCOMO-II as base model. The experimental results show that our method could significantly improve prediction accuracy of conventional Functional Link Artificial Neural Networks (FLANN) and has potential to become an effective method for software cost estimation.

Keywords: Software cost estimation, Genetic algorithm, FLANN, COCOMO-II.

1 Introduction

Software cost estimation is critical for the success of software project management. It affects management activities including resource allocation, project bidding and project planning. The importance of accurate estimation has led to extensive research efforts to software cost estimation methods. From a comprehensive review, these methods could be classified into following six categories: parametric models including COCOMO, SLIM and SEER-SEM, expert judgment including Delphi technique and work break down structure based methods, learning oriented techniques

including machine learning methods and analogy based estimation; regression based methods including ordinary least square regression and robust regression; dynamics based model, composite methods. [3]

In this paper, we are concerned with cost estimation models that are based on Functional Link artificial neural networks .The Functional Link artificial neural network (FLANN) architecture for predicting software development effort is a single-layer feed forward neural network consisting of one input layer and an output layer. The FLANN generates output (effort) by expanding the initial inputs (cost drivers) and then processing to the final output layer. Each input neuron corresponds to a component of an input vector. The output layer consists of one output neuron that computes the software development effort as a linear weighted sum of the outputs of the input Layer [9]. The large and non-normal data sets always lead FLANN methods to low prediction accuracy and high computational complexity. To alleviate these drawbacks our proposed idea has been devoted to simultaneously optimize selected class of projects and their feature selection by genetic algorithm (GA).

GA is used to optimize the selected class of projects and their feature weights. Functional Link Neural networks and cost estimation fundamentals are briefly reviewed in section 2. The proposed GA approach for optimizing the selected class of projects is described in section 3. In section 4, numerical examples from COCOMO dataset is used to illustrate the performance. A conclusion and overview of future work conclude this paper.

2 Background

2.1 The COCOMO

The Constructive Cost Model, COCOMO, was introduced by Boehm [2]. It has become one of the most widely used software cost estimation models in the industry. To support new life cycles and capability, it has evolved into a more comprehensive estimation model, called COCOMO II [1]. COCOMO II consists of three sub models, each one offering increased fidelity. Listed in increased fidelity, these sub models are called Application Composition, Early design and Post Architecture models. Until recently, only the last and most detailed sub model, Post Architecture had been implemented in a calibrated software tool. As such, unless otherwise explicitly specified, all further references in this study to COCOMO II can be assumed to be the Post Architecture Model [5].

2.2 Architecture of FLANN [9,10]

The FLANN network can be used not only for functional approximation but also for decreasing the computational complexity. This method is mainly focused on functional approximation. In the aspect of learning, the FLANN network is much faster than other network. The primary reason for this is that the learning process in FLANN network has two stages and both stages can be made efficient by appropriate learning algorithms. The use of on FLANN to estimate software development effort requires the determination of its architecture parameters according to the characteristics of COCOMO.

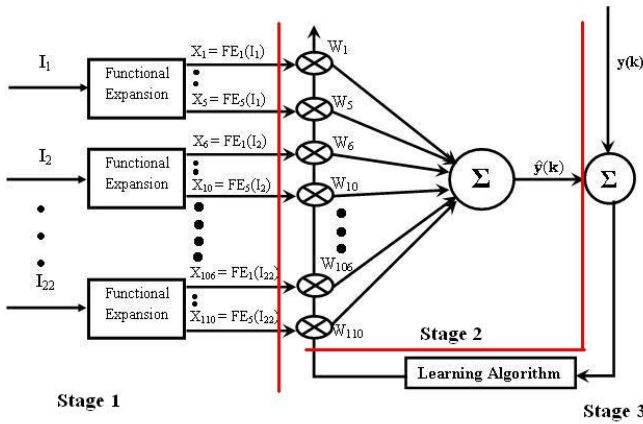


Fig. 1. FLANN Architecture

A general structure of FLANN is shown in figure 1. FLANN is a single-layer nonlinear network. Let k be the number of input-output pattern pairs to be learned by the FLANN. Let the input pattern vector X_k be of dimension n , and the output y_k be a scalar. The training patterns are denoted by $\{X_k, y_k\}$. A set of N basis functions $\Phi(X_k) = [\phi(X_k) \ \phi(X_k) \ \dots \ \phi(X_k)]^T$ are adopted to expand functionally the input signal $X_k = [x_1(k) \ x_2(k) \ \dots \ x_n(k)]^T$. These N linearly independent functions map the n -dimensional space into an N -dimensional space, that is $R^n \rightarrow R^N$, $n < N$.

The linear combination of these function values can be presented in its matrix form, that is $S = W\Phi$. Here $S_k = [S_1(k) \ S_2(k) \ \dots \ S_m(k)]^T$, W is the $m \times N$ dimensional weight matrix. The matrix S_k is input into a set of nonlinear function $\rho(\bullet) = \tanh(\bullet)$ to generate the equalized output $\hat{Y} = [\hat{y}_1 \ \hat{y}_2 \ \dots \ \hat{y}_m]^T$, $\hat{y}_j = \rho(S_j)$, $j = 1, 2, \dots, m$.

STAGE-1: The 22 cost factors of the validated dataset are taken as the input of the network. These factors are then expanded functionally by using the following formulas.

Chebyshev polynomials are given by:

$$\begin{aligned} T_0(x) &= 1, & T_1(x) &= x, & T_2(x) &= 2x^2 - 1, \\ T_3(x) &= 4x^3 - 3x, & T_4(x) &= 8x^3 - 8x^2 + 1. \end{aligned}$$

The Chebyshev Polynomial FLANN(C-FLANN) is implemented and comparative study with conventional methods is shown in Table 1. Here after C-FLANN will be expressed as FLANN.

STAGE-2: Let each element of the input pattern before expansion be represented as $z(i)$, $1 < i < I$ where each element $z(i)$ is functionally expanded as $z_n(i)$, $1 < n < N$, where N = number of expanded points for each input element. In this study, $N = 5$ and I = total number of features in the dataset has been taken.

Expansion of each input pattern is done as follows:

$$\begin{aligned} x_0(z(i)) &= 1, & x_1(z(i)) &= z(i), \\ x_2(z(i)) &= 2z(i)^2 - 1, & x_3(z(i)) &= 4z(i)^3 - 3z(i), & x_4(z(i)) &= 8z(i)^3 - 8z(i)^2 + 1. \end{aligned}$$

where, $z(i)$, $1 < i < d$, d is the set of features in the dataset.

These nonlinear outputs are multiplied by a set of random initialized weights from the range [-0.5, 0.5] and then summed to produce the estimated output $y(k)$. All the $Y(k)$'s are summed to get $\hat{y}(k)$.

STAGE 3: TRAINING DATA

This output is compared with the corresponding desired output and the resultant error for the given pattern is used to compute the change in weight in each signal path P , given by

$$\Delta W_j(k) = \mu \times x f_j(k) \times e(k)$$

where, $x f_j(k)$ is the functionally expanded input at k^{th} iteration.

If there are p patterns to be applied then average change in each weight is given by

$$\overline{\Delta W_j(k)} = \frac{1}{p} \sum_{i=1}^p \Delta W_j^i(k)$$

Then the equation, which is used for weight update, is given by

$$W_j(k+1) = W_j(k) + \Delta W_j(k)$$

where, $W_j(k)$ is the j^{th} weight at the k^{th} iteration, μ is the convergence coefficient, its value lies between 0 to 1 and $1 < j < J$, $J = M \times d$. M is defined as the number of functional expansion unit for one element.

$$e(k) = y(k) - \hat{y}(k)$$

where, $y(k)$ is the target output and $\hat{y}(k)$ is the estimated output for the respective pattern and is defined as:

$$\hat{y}(k) = \sum_{j=1}^J x f_j(k) \cdot w_j(k)$$

where, $x f_j$ is the functionally expanded input at k^{th} iteration and $W_j(k)$ is the j^{th} weight at the k^{th} iteration and $W_j(0)$ is initialized with some random value from the range [-0.5, 0.5].

3 Framework

Genetic algorithm (GA) is a stochastic global optimization technique initially introduced by Holland in 1970's [4]. By mimicking biological selection and reproduction, GA can efficiently search through the solution space of complex problems and it is naturally parallel and provides opportunity to escape from local optimum. GA has become one of the most popular algorithms for optimization problems. In this section, we construct the OCFWFLANN system (stands for Optimal projects of predicted Class and Feature Weighting and Functional Link Artificial Neural Network based Estimation) which can perform simultaneous optimization of 'N' projects of the predicted class and their feature weights. GA is selected as

optimization tool for OCFWFLANN system. The detailed description is presented in section 3.2. In order to introduce the fitness function in GA algorithm, performance metrics for estimation accuracy are firstly presented in the section 3.1.

3.1 Performance Evaluation Metrics

To measure the accuracies of the proposed methods, three performance metrics are considered: Mean Magnitude of Relative Error (MMRE), Median Magnitude of relative error (MdMRE), and PRED (0.25), because these measures are widely accepted in literature [4].

The MMRE is defined as:

$$MMRE = \frac{1}{n} \sum_{i=1}^n MRE \quad (1)$$

$$MRE = \left| \frac{E_i - \hat{E}_i}{E_i} \right| \quad (2)$$

Where n denotes the total number of projects, E_i denotes the actual cost of ith project, and \hat{E}_i denotes the estimated cost of the ith project. Small MMRE value indicates the low level of estimation error. However this metric is unbalanced and penalizes overestimation more than underestimation. The MdMRE is the median of all the MREs.

$$MdMRE = \text{Median} (MRE) \quad (3)$$

It exhibits similar pattern to MMRE but it is more likely to select the true model especially in the underestimation case since it is less sensitive to extreme outlier [6]. The PRED (0.25) is the percentage of prediction that fall within 25 percent of actual cost

$$PRED(q) = \frac{k}{n} \quad (4)$$

Where n denotes the total number of projects and k denotes the number of projects whose MRE is less than or equal to q. normally, q is set to be 0.25. The PRED (0.25) identifies cost estimations that are generally accurate, while MMRE is biased and not always reliable as a performance metric. However, MMRE has been de facto standard in the software cost estimation literature. Thus MMRE is selected for the fitness function in GA. More specifically, for each combination of 'N' project parameters and cost driver weights, MMRE is computed across the validation dataset. Then GA searches through the project parameter space to minimize MMRE.

3.2 GA for Optimization

The procedure for OPFWFLANN system via Genetic Algorithm is presented in this section. The system consists of two stages: the first one is training stage and the second one is testing stage. In the training stage 93 NASA data points are presented to the system, the FLANN is configured with cost drivers to produce the cost prediction for the given input data point. The class label of the input data is determined basing on PM obtained. GA explores the class space to minimize the error (in terms of MMRE) of FLANN on the training projects by the following steps:

3.2.1 Encoding

To apply GA for searching, the cost drivers are encoded as binary string chromosome. Each individual chromosome consists of number of binary digits. There are six features for each driver (very low-vl, low-l, nominal-n, high-h, very high-vh, extra high-xh). Here we encode cost driver weights with 3 bits (0-n, 1-vl,2-l,3-n,4-h,5-vh,6-xh,7-n). 0 and 7 are assumed default values nominal (n). Since the cost driver weights are decimal values, before entering into FLANN model these binary codes are transformed into decimal numbers.

3.2.1.1 Population Generation and Fitness Function. After encoding the individual chromosome, the system then generates a population of chromosomes. Each chromosome is evaluated by the fitness function in GA. Since the GA is designed to maximize the fitness value and the smaller MMRE value indicates more accurate prediction, we set the fitness function as the reciprocal of MMRE.

$$f = \frac{1}{MMRE} \quad (5)$$

Given one training project as input, ANN predicts the PM for the project, basing on the person month; the class is identified for the project, which contains set of similar projects as input. To evaluate the prediction performance of the ANN model, the error metric MMRE, PRED (0.25), and MdMRE applied on the training project set in the class. Then, the reciprocal of MMRE is used as the fitness value for each cost driver combination (or chromosome).

3.2.2 Rules for Selection, Extinction and Multiplication

The standard roulette wheel is used to select chromosomes from the current population. The selected chromosome were consecutively paired with a probability of 0.8 was used to produce new chromosome in each pair. The newly created chromosome constituted a new population. The population is evolved by GA algorithm using evolutionary rules described above. The individual with best fitness value is in the population in every cycle.

3.2.3 Completion of Evaluation

The population is evolved by the GA algorithm in the first stage until the number of generations is equal to or excess 2000 or the best fitness value did not change in the last 200 generations. The second stage is the testing stage. In this stage the system receives the optimized parameters from the training stage to configure the FLANN model. The optimal FLANN is then applied to the testing project to evaluate the performance of the trained FLANN.

4 Experimentations and Results

The COCOMO NASA 2 Dataset containing 93 data points have been taken for our experiment [6]. The data is in COCOMO-I format calibrated to COCOMO-II using Rosetta stone [1]. COCOMO measures efforts in the calendar months of 152 hours (and includes development and management hours). COCOMO assumes that the

effort grows more than linearly on software size. There are total of 17 effort multipliers in COCOMO II. These cost factors are expressed in 6 forms i.e v_l, l, n, h, v_h, h_x . For our experimental results we have included 5 scaling factors and assumed their values as “nominal”. Along with these there are two more attributes, namely, KSLOC (Kilo Source Lines of Code) and actual effort. There are 11 classes distributed across 93 data points .

For the purpose of validation, we adopt three-fold cross validation [4] to evaluate accuracy of the methods. In this scheme the NASA dataset is randomly divided into three nearly equal sized subsets. At each time one of three subsets is used as the test sets which is exclusively used to evaluate the estimation performance, and other two subsets treated as Validation data set and training data set exclusively used to optimize the cost drivers. This process is repeated three times. Then the average training error and testing error across all three trials are computed. The advantage of this scheme is it matters less how the data is split since each data point is assigned into a test set, a training set and a validation set respectively once. In the experiment we apply three types of FLANN based models. The first model is conventional FLANN [9,10], the second model is OFWFLANN (GA optimizing feature weights(cost drivers) for Functional Link Neural Network based cost estimation) which does not optimize the projects data points in the corresponding class. The third model is OCFWFLANN which uses GA simultaneously optimize the class and the feature weights (cost drivers).For comparison, other popular estimation models including Step Wise Regression (SWR) [7], Classification and Regression Trees (CART) [8], are also included in the experiments.

The experimental results are summarized in table 1. It shows that OCFWFLANN achieves the best level of prediction performance (0.27 for MMRE, 0.19 for MdMRE, and 0.26 for PRED (0.25)).

Table 1. Results and Comparison on NASA Dataset

Methods	MMRE		MdMRE		PRED(0.25)	
	Training	Testing	Training	Testing	Training	Testing
OCFWFLANN	0.28	0.27	0.24	0.19	0.31	0.26
OFWFLANN	0.38	0.33	0.39	0.28	0.30	0.38
FLANN	0.43	0.37	0.37	0.33	0.46	0.39
SWR	0.92	0.79	0.58	0.44	0.45	0.41
CART	0.85	0.64	0.48	0.37	0.34	0.30

5 Conclusion and Future Work

On appraising the above novel technique the hybrid system of GA and FLANN provides better prediction accuracy compared to FLANN. GA is used as a tool for simultaneously optimizing the concerned class to which the input project belongs and cost drivers. The experimental results show that our method gives pacifying optimal

performance as compared to conventional FLANN and outperform the comparative techniques such as OFWFLANN, SWR and CART. Motivation is therefore exploring the scope of application of soft computing in the field of Software Cost Estimation.

We have done our research in the direction of software cost estimation by hybrid system using GA as it has not been explored extensively till date. There are numerous cost estimation techniques that have been proposed in different real-world applications. We extend connotations to our work with Artificial Bee Colony (ABC), Differential Evolution (DE), Artificial Immune System (AIS), Bacterial foraging optimization algorithm, Neuro Fuzzy, Neuro Genetic, Simulated Annealing and fuzzy logic.

References

1. Boehm, B., Abts, C., Brown, A., Chulani, S., Clark, B., Horowitz, E., Madach, R., Reifer, D., Steece, B.: *Software Cost Estimation with COCOMO II*. Prentice Hall, Upper Saddle River (2000)
2. Boehm, B.: *Software Engineering Economics*. Prentice Hall (1981)
3. Li, Y.F., Xie, M., Goh, T.N.: A study Of Project Selection Feature Weighting For Analogy Based Software Cost Estimation. *The Journal of Systems and Software* 82, 241–252 (2009)
4. Li, Y.F., Xie, M., Goh, T.N.: Optimization of Feature Weights and Number of Neighbors For Analogy Based Cost Estimation in Software Project Management. In: *Proceedings of the 2008 IEEE IEEM (2008)* ISBN: 978-1-4244-2630-0/08
5. Musilek, P., Pedrycz, W., Sun, N.: On the Sensitivity of COCOMO II Software Cost Estimation Model. In: *METRICS 2002, The Proceedings of 8th IEEE Symposium on Software Metrics (2002)* ISBN 0-7695-1339-5/02
6. Menzies, T.: *The PROMISE Repository Of Software Engineering Databases*. In: *School of Information Technology and Engineering, University of Ottawa, Canada (2006)*, <http://promise.site.uottawa.ca/SERepository>
7. Shepperd, M., Kadoda, G.: Comparing Software Prediction Techniques using Simulation. *IEEE Transaction on Software Engineering* 27(11), 1014–1022 (2001)
8. Stensrud, E.: Alternative Approaches to Software Prediction of ERP Projects. *Information and Software Technology* 43(7), 413–423 (2001)
9. Tirimula Rao, B., Sameet, B., Kiran Swathi, G., Vikram Gupta, K., Raviteja, C., Sumana, S.: A Novel Neural Network approach for Software Cost Estimation Using Functional Link Artificial Neural Networks. *International Journal of Computer Science and Network Security (IJCSNS)* 9(6), 126–131 (2009)
10. Zhao, H., Zeng, X., Zhang, J., Li, T., Liu, Y., Ruan, D.: Pipelined functional link artificial recurrent neural network with the decision feedback structure for nonlinear channel equalization. *Information Sciences* 181(17), 3677–3692 (2011)

Detecting and Searching System for Event on Internet Blog Data Using Cluster Mining Algorithm

Robin Singh Bhadoria¹, Manish Dixit², Rohit Bansal³, and Abhishek Singh Chauhan⁴

¹ Dept of Computer Science & Engineering, IITM, Gwalior (MP)

ssr_robin@yahoo.co.in

² Dept of Computer Science & Engineering, MITS, Gwalior (MP)

dixitmits@gmail.com

³ NIIST, Bhopal(MP)

rohitbansal.cse@gmail.com

⁴ SATI, Vidisha(MP)

abhichauhan78@gmail.com

Abstract. The popularity of Internet is growing every day with an exponential growth in the information that is being published over it. Apart from static content, dynamic content on the Web is also growing at an increasing rate thanks to blogs, news forums and the likes. Users of such blogs and forums write about their personal life, professional life and events happening in real world such as a cricket match, elections, a product release or disasters. The number of blog entries published on an event is proportional to its popularity. Using this as the basis, we designed a system called EventDS (Event Detection and Searching) which detects major events by analyzing blogs using a novel clustering algorithm called PDDPHAC. We also propose a new representation for events: each event is represented as a Topic Tree where sub-topics are treated as children of their super-topics.

1 Introduction

With advances in technology today Internet is easily available to most of the people. Apart from all other activities that happen in Internet, a lot people express their feelings, views and share their knowledge over the Internet. They use blogs, social networking sites and discussion forums to express and discuss their opinions as these are available to everyone for free and provide unlimited space. Now people don't have to host their own websites to express something, they can use these mediums and hence the information that is being published on the Internet through these mediums is growing at a very fast rate. It has been reported that more than 112 million blogs exist today and more than 175,000 new blogs are created everyday and more than 1.6 millions posts are created per day [23].

2 Algorithm for System

2.1 Clustering Algorithms

In this paper, we have studied and used two well know clustering algorithms PDDP (Principal Direction Divisive Partition) and HAC (Hierarchical Agglomerative

Clustering). Both are hierarchical based clustering algorithms but PDDP is a top-down clustering algorithm while HAC is a bottom-up clustering algorithm. We discuss them in the following subsections in detail.

2.1.1 Principal Direction Divisive Partition (PDDP)

PDDP is hierarchical divisive partitioning clustering algorithm was introduced by Boley [5]. As mentioned earlier PDDP is top-down clustering algorithm so it first considers the entire document set as a single cluster and splits it into two clusters and recursively selects a cluster with maximum distortion and splits into further into two clusters until a terminating condition is reached. PDDP forms binary tree of clusters where root of the tree contains all documents and leaf nodes are the final output clusters.

2.1.2 Hierarchical Agglomerative Clustering (HAC)

HAC is a hierarchical bottom-up clustering [24] and it is the one of the most highly studied clustering algorithms as it is very intuitive. It first treats every document in the document set as a cluster and merges the two most similar clusters in each iteration. It uses document-document distance matrix to combine two most similar clusters and updates the distance matrix after each iteration. HAC updates the distance matrix in different ways depending on linkage. The most well known and used linkages are discussed below.

3 Event Representation, Tracking and Updater

Now event identification is trivial, we consider clusters which have the number of documents more than EventTh as event clusters. Although we have designed PDDPHAC in such a way that it produces better event detection results than PDDP and HAC algorithms the performance of PDDPHAC may depend on the performance of individual algorithms to some extent. That is its results can sometimes get effected by the threshold used in HAC to combine clusters returned by PDDP. To obtain reliable results, we have used a tight threshold for HAC which results compact clusters so some events may not be detected.

3.1 Event Representation

The main goal for representing an event in a topic tree is to capture different topics being discussed part of the event. We have used the following idea to capture the topics in an event. First we treat an event document as an entire dataset and recomputed TF-IDF scores of terms in them, and then we apply PDDPHAC on this event document set and get the topic clusters. We then use the topic clusters to construct topic tree of the event.

3.1.1 Topic Tree Structure

Topic tree for an event is constructed using the following steps: a node's children are subtopics in it. In a topic tree, the root of tree contains information about all the documents of the event, and each child of the root represents a subtopic of it. All the subtopics of a node are not equally important and some of them can contain a larger

number of documents compared to other nodes, so we call them prominent topics. And these topics can further contain subtopics so we again apply PDDPHAC on the prominent topics and get subtopics. In this way we proceed until there are no prominent topics.

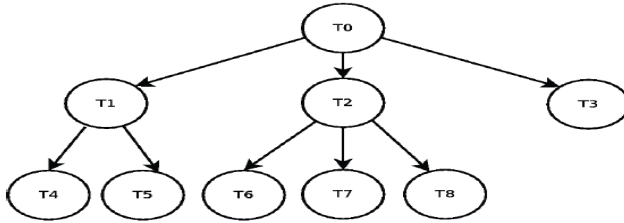


Fig. 1. Topic Tree

Figure 1 shows the sample topic tree, where T0 is the root of the tree so it contains all the information about the event. And T1, T2, T3 are subtopics of T0 so they are more specific than T0 and further T4 and T5 are subtopics of T1 and T6, T7 and T8 are subtopics of T2. The leaf nodes: T4, T5, T6, T7 and T8 are the most specific topics in the topic tree.

3.2 Event Tracking

Event tracking implies identifying newly written blogs which are related to the already detected events. The event tracking can be done in following ways:

3.3 Event Updater

We now describe possible scenarios of updates and in all cases we make sure that the key property of topic tree mentioned earlier does not get violated.

1. Updating a very similar topic: This scenario arises when we are updating a topic node with newly posted blogs related to the topic. And sometimes newly tracked blogs contain very similar information, in this case creating new topic node is not necessary, we just update existing topic node. The following Figure 2 explains this case.

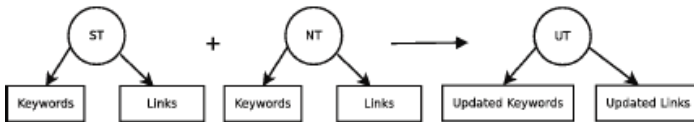


Fig. 2. Describes Updating a very similar topic

where ST is a similar topic, NT is a new topic and UT is an updated topic. Keyword set is updated by recomposing overall principal direct vector and ranking keywords according to it. Link set is also updated by taking top 1 blog posts from both of the link sets of ST and NT.

2. Updating a leaf topic: This case occurs when we have tracked sufficient documents related to a leaf topic node of topic tree and a new topic created from these documents does not satisfy previous update condition. We create a new topic node using the tracked documents and update the corresponding related node in the following manner.

where LT is a leaf topic, NT is a new topic and NPT is a new parent topic. In this case the reason for creating a new topic node is to ensure the key property of the topic tree. If we create NT as single child of LT then this property will be violated as discussed earlier. NT is not very similar to LT if that was the case then it should have come under first case. So in this case, we update the topic tree by creating a new parent node NPT of both LT and NT, and keyword and link sets of NPT are computed by combining keyword and link sets of both LT and NT respectively.

3. Updating a topic node other than leaf node: This case occurs when the both the above update conditions are not satisfied. In this case we have to update a node which have at least 2 children. This update is performed in the following way.

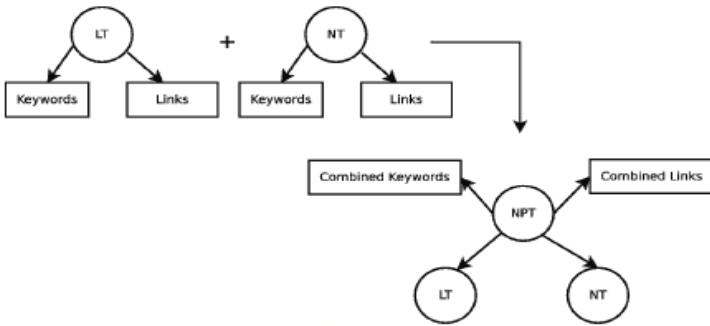


Fig. 3. Describes Updating a leaf topic

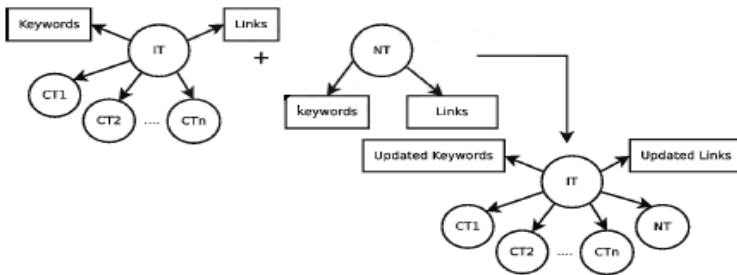


Fig. 4. Describes Updating a topic node which have children

4 Result and Analysis

4.1 PDDP Analysis

In this work, we discuss the experiments conducted on PDDP to test its performance on our dataset. The Table 2 presents the results produced by PDDP using CSV as the termination condition.

Table 1. PDDP clustering results

<u>Parameters</u>	<u>Values</u>
Actual Events	5
Detected Events	1
Number of clusters	12
Purity of clusters	90.74%
Purity of event clusters	100%
Execution time (in seconds)	0.0005
Clustered Documents	303
Remaining Documents	2197

This experiment was to understand PDDP splitting, the Figure 5 shows, how the dataset was partitioned into two clusters after first iteration. PDDP puts documents with a projection value ≥ 0 in one cluster and the rest of them in the other cluster. According to this, in Figure 5, we can see that PDDP partitioned three events namely Haiti Earthquake, Sachin Double Century, Iceland Volcano properly, but the other two: Avatar Movie release and Indian Budget 2010 are improperly partitioned. As we discussed earlier, PDDP doesn't perform well on ill-separated datasets and the figure shows the same.

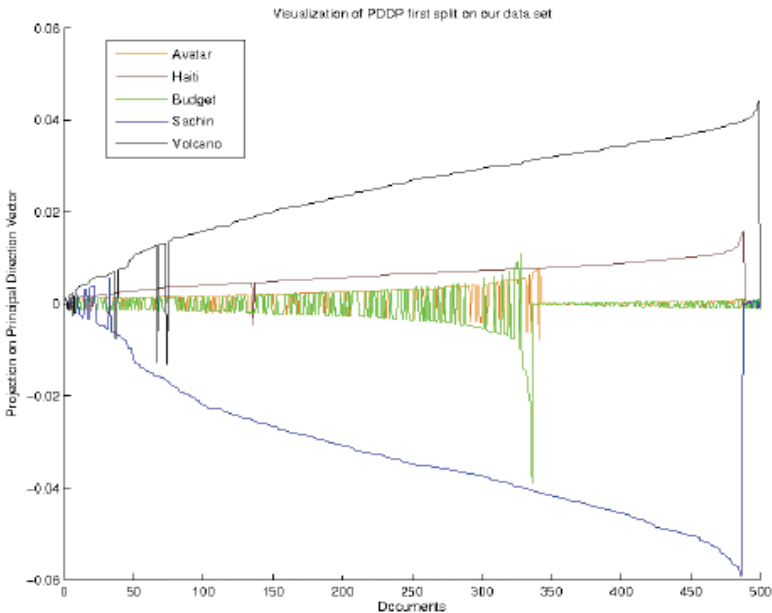


Fig. 5. Projection values of the documents in PDDP first iteration

Table 2, we present details about the same experiment but now we show the number of documents from each event class in each cluster.

Table 2. Data Distribution after first PDDP split

Event name	Number of blogs	
	Cluster ₁	Cluster ₂
Avatar Movie Release	358	142
Haiti Earthquake	489	11
Indian Budget 2010	207	291
Sachin Double Century	21	479
Iceland Volcano	493	7

HAC Analysis

In this section, we discuss the experiments that have been conducted to test HAC on our dataset.

Table 3. Shows the results produced by HAC on our test dataset

Parameters	Value		
	DT=0.5	DT=0.6	DT=0.7
Actual Events	5	5	5
Detected Events	0	2	1
Number of clusters	1746	1154	531
Purity of clusters	99.32%	85.48%	40.56%
Purity of event Clusters	-----	62.05%	21.79%
Execution time(in sec.)	0.1751	0.2420	0.2454
Clustered Documents	0	917	1886
Remaining Documents	250	1583	614

PDDPHAC Analysis

In previous sections, we have shown the experiments conducted for testing PDDP and HAC clustering algorithms on our test dataset. As we can see PDDP had performed better than HAC on our dataset but its accuracy of event detection was poor.

Table 4. PDDPHAC clustering results

Parameters	Values			
	CS = 5	CS = 10	CS = 15	CS = 20
Actual Events	5	5	5	5
Detected Events	5	5	5	5
Number of clusters	114	70	43	29
Purity of clusters	94.16%	93.88%	93.32%	93.04%
Purity of event clusters	98.35%	98.71%	98.19%	97.48%
Execution time (in seconds)	0.0054	0.0041	0.0022	0.0018
Clustered Documents	2000	2089	2150	2223
Remaining Documents	500	411	350	277

where CS is cluster size and HACDT is HAC distance threshold

From Table 4, we can easily say that the performance of PDDPHAC is much better than both PDDP and HAC. The event detection accuracy of PDDPHAC is more than the others and the time taken by PDDPHAC for clustering is comparable to PDDP and much less than HAC.

4.2 Architecture of EventDS

We have discussed the main parts of the EventDS: Back-end and Front-end, we now present the architecture of complete EventDS in Figure 6. The figure shows architecture of both Back-end and Front-end and how they are connected using a MySQL database. The figure shows all the elements of EventDS and their integration. Observe the arrows between Back-end and Database, they are bidirectional i.e. Back-end reads data from and writes data to the database. But arrows between Front-end and Database are unidirectional i.e. Front-end just reads data from the database but never modifies data in the database. From the figure, we can observe that the tasks performed by the Back-end are more complex and major compared to the Front-end. The Front-end just provides an interface to access the functionalities provided in the Back-end.

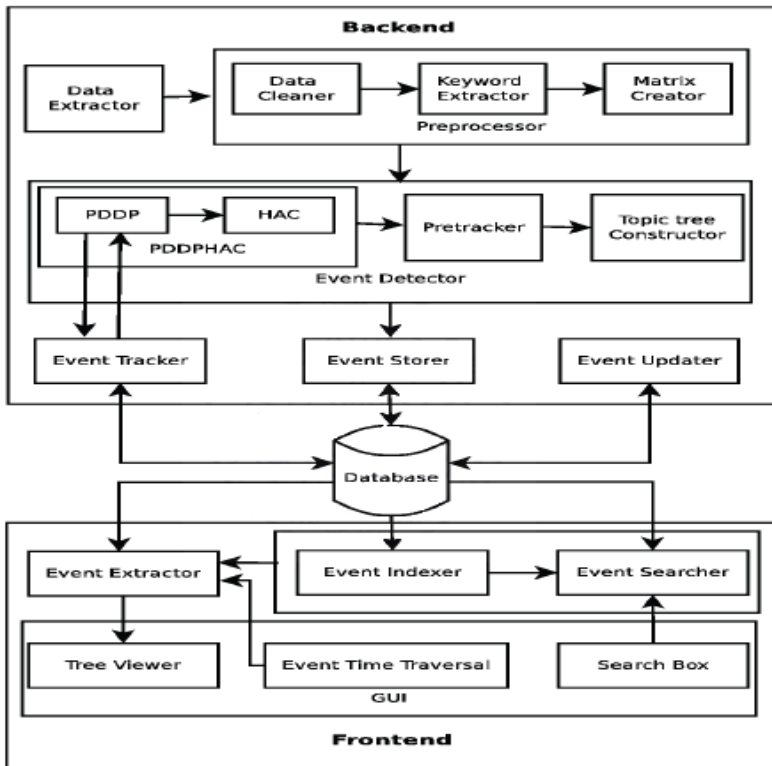


Fig. 6. Architecture of EventDS

5 Conclusion

We have studied two very famous clustering algorithms called Principal direction divisive partitioning (PDDP) and Hierarchical agglomerative clustering (HAC) and tested their applicability in detecting events from blogs. The results are not very satisfactory so we designed a new clustering algorithm called PDDPHAC by combining these two clustering algorithms and proved that PDDPHAC outperformed both algorithms. Using the proposed clustering algorithm, we have designed a system called EventDS: events detection and searching and the results produced by EventDS were very satisfactory. We have also proposed a new representation for the events that is topic tree and discussed its advantages over using weighted keyword set representation. We have implemented event tracker for tracking the newly written blogs on already detected events and event updater for updating the events.

References

- [1] Aksyonof, A.: Sphinx: free open-source SQL full-text search engine, <http://sphinxsearch.com/>
- [2] Trevor, H., Robert, T., Jerome, F.: Hierarchical clustering. In: *The Elements of Statistical Learning*, 2nd edn., pp. 520–528. Springer, Heidelberg (2009)
- [3] Belmonte, N.G.: JIT: Java Script InfoVis Toolkit, <http://thejit.org/>
- [4] Boley, D.: Principal Direction Divisive Partitioning. *Data Mining and Knowledge Discovery* 2, 325–344 (1997)
- [5] Boley, D.: Hierarchical Taxonomies using Divisive Partitioning. Technical report (1998)
- [6] Boley, D., Borst, V.: Unsupervised Clustering: A Fast Scalable Method for Large Datasets. Technical report (1999)
- [7] Thierer, A.: Internet Statistics, <http://techliberation.com/2008/05/06/need-help-how-many-blogs-are-there-out-there/>
- [8] Denton, N.: Weblogs: Blogspot updates provider, <http://www.weblogs.com/>
- [9] Fiscus, J.G., Doddington, G.R.: Topic detection and tracking evaluation overview. In: *Topic Detection and Tracking: Event-Based Information Organization*, pp. 17–31. Kluwer Academic Publishers (2002)
- [10] Manning, C.D., Raghavan, P., Schtze, H.: Evaluation of clustering. In: *Introduction to Information Retrieval*, pp. 356–360. Cambridge University Press (2008)
- [11] Witten Ian, H., Paynter Gordon, W., Eibe, F., Carl, G., NevillManning Craig, G.: KEA: practical automatic keyphrase extraction. In: *DL 1999: Proceedings of the Fourth ACM Conference on Digital Libraries*, pp. 254–255. ACM, New York (1999)
- [12] Seo Katia, Y.-W., Seo, Y.W., Sycara, K.: Text Clustering for Topic Detection. Technical report (2004)
- [13] Kruengkrai, C.: Implementation of PDDP Algorithm in JAVA, <http://www.tcllab.org/canasai/software/omniclusterer>
- [14] Kruengkrai, C., Sornlertlamvanich, V., Isahara, H.: Document Clustering Using Linear Partitioning Hyperplanes and Reallocation. In: Myaeng, S.-H., Zhou, M., Wong, K.-F., Zhang, H.-J. (eds.) *AIRS 2004. LNCS*, vol. 3411, pp. 36–47. Springer, Heidelberg (2005)
- [15] Lenz, H.J.: Proximities in Statistics: Similarity and Distance. In: *Preferences and Similarities*. CISM International Centre for Mechanical Sciences, vol. 504, pp. 161–177. Springer, Vienna (2008)

- [16] Zhang, K., Xu, H., Tang, J., Li, J.: Keyword Extraction Using Support Vector Machine. In: Yu, J.X., Kitsuregawa, M., Leong, H.-V. (eds.) WAIM 2006. LNCS, vol. 4016, pp. 85–96. Springer, Heidelberg (2006)
- [17] Matsuo, Y., Ishizuka, M.: Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information (2003)
- [18] Trevor, H., Robert, T., Jerome, F.: Hierarchical clustering. In: *The Elements of Statistical Learning*, 2nd edn., pp. 520–528. Springer, Heidelberg (2009)
- [19] Palshikar, G.K.: Keyword Extraction from a Single Document Using Centrality Measures. In: Ghosh, A., De, R.K., Pal, S.K. (eds.) PReMI 2007. LNCS, vol. 4815, pp. 503–510. Springer, Heidelberg (2007)

Knowledge Based Evolutionary Programming: Cultural Algorithm Approach for Constrained Optimization

Bidishna Bhattacharya^{1,*}, Kamal Mandal², and Niladri Chakraborty²

¹ Techno India, Saltlake, Electrical Engineering Dept., India
bidishna_inf@yahoo.co.in

² Jadavpur University, Department of Power Engineering, India

Abstract. A cultural approach to solve the problem defined by the economic load dispatch in power systems is presented in this paper. The practical problems of economic load dispatch have non-smooth cost functions with equality and inequality constraints that make the problem of finding the global optimum difficult using any mathematical approaches. Our approach is based on the concept of a cultural algorithm and is applied to constrained optimization problems in which a map of the feasible region is used to guide the search more efficiently. It combines cultural algorithm with evolutionary programming technique in such a way that a simple evolutionary programming (EP) is applied as a based level search, which can give a good direction to the optimal global region, and a domain knowledge (using the concept of cultural algorithm) is used as a fine tuning to determine the optimal solution at the final. The effectiveness and feasibility of the proposed method is tested on a practical thirteen generator system. Results obtained by the proposed method are compared with the other evolutionary methods. It is seen that the proposed method can produce comparable results.

1 Introduction

Economic load dispatch is one of the important tasks in the operation of power systems. The main objective is allocate the generation of the committed units in such a manner that overall operating cost is minimum satisfying a set of linear and non-linear constraints. Previous efforts on solving ELD problems have employed various conventional methods like lambda iteration, quadratically constrained programming, gradient methods etc [18], [1], [10]. These early methods were unable to meet the exact requirement leaving some approximate values which are basically not optimal value & hence a huge revenue loss occurs for nonlinear characteristics in practical systems due to valve point loading, prohibited operating zones and ramp rate limits of generators. After evolution of artificial intelligence technique, several population-based optimization methods are used to play the important role to solve the problem of economic load dispatch.

In the past decade, the global optimization techniques like genetic algorithms (GA) [4], particle swarm optimization (PSO) [8], evolutionary programming [9] etc which

* Corresponding author.

are form of probabilistic heuristic algorithm, has been successfully used to solve economic load dispatch problems.

Reynolds adds a new technique Cultural Algorithm as a vehicle for modeling social evolution and learning the behavioral traits [13]. Cultural algorithm is basically a global optimization technique which consists of an evolutionary population space whose experiences are integrated into a Belief space which influences the search process to converge the problem in a direct way. Cultural algorithms have been successfully applied to global optimization of constrained problems [14].

In this paper, an alternative approach to cultural algorithm has been proposed to solve the ELD problems. We worked on cultural algorithm embedded in evolutionary programming to solve the economic dispatch problem involving valve point loading effect. Embedding an EP into a CA framework [6] was developed to investigate the influence of global knowledge on the solution of optimization problem and it is successful to solving non-constrained optimization problem in previous work [5]. In this paper we use EP embedded in CA to solve constrained optimization problem.

2 Problem Formulation

The pure Economic Load Dispatch (ELD) problem is one of the major problems in power system operation and planning. The classical ELD problem may be described by minimizing the total fuel cost of the generating units under several operating constraints. The fuel cost curve for any unit is assumed to be approximated by segments of quadratic functions of the active power output of the generator. For a given power system network, the problem may be described as optimization (minimization) of total fuel cost as defined by (1) under a set of operating constraints.

$$FC(P_g) = \sum_{i=1}^n (a_i P_i^2 + b_i P_i + c_i) \quad (1)$$

where $FC(P_g)$ is the total fuel cost of generation in the system (\$/hr), a_i, b_i, c_i are fuel cost coefficients of the i th generating unit, P_i is power generated by the i th unit, and n is the number of thermal units. The coefficients a_i, b_i and c_i are generally obtained by curve fitting.

However, for more practical and accurate modeling of fuel cost function, the above expression is to be modified suitably. Modern thermal power plants consist of generating units having multi-valve steam turbines in order to incorporate flexible operational facilities. The generating units with multi-valve turbines have very different cost curve compared with that defined by (1) and exhibit a greater variation in the fuel cost curves. Typically, ripples are introduced in the fuel cost curve as each steam valve starts to operate. The valve-point effect may be considered by adding a sinusoidal function [16] to the quadratic cost function described above. Hence, the problem described by (1) is revised as follows:

$$FC_v(P_g) = \sum_{i=1}^n (a_i P_i^2 + b_i P_i + c_i + |e_i \times \sin(f_i \times (P_{i,\min} - P_i))|) \quad (2)$$

where $FC_v(P_g)$ is total fuel cost of generation in (\$/hr) including valve point loading, e_i, f_i are fuel cost coefficients of the i th generating unit reflecting valve-point effect.

The cost is minimized with the following generator capacities and active power balance constraints as:

$$P_{i,\min} \leq P_i \leq P_{i,\max} \quad (3)$$

$$\sum_{i=1}^n P_i = P_D + P_L \quad (4)$$

where, $P_{i,\min}$ and $P_{i,\max}$ are the minimum and maximum power generation by i th unit respectively, P_D is the total power demand and P_L is total transmission loss.

The transmission loss P_L can be calculated by using B matrix technique and is defined by (5) as

$$P_L = \sum_{i=1}^n \sum_{j=1}^n P_i B_{ij} P_j \quad (5)$$

where B_{ij} 's are the elements of loss coefficient matrix B.

3 Cultural Algorithm (CA)

The basic concept, principles & mechanisms of every evolutionary technique are based on how natural systems evolve to solve the complex computational problems. The CA works on the concept that, in advance societies the individuals get improvement by not only the information which it possesses due to heredity but by the information which are acquired after years of experience, which is called culture. This cultural evolution is an inheritance process that operates at two levels: the micro evolutionary level and the macro evolutionary level [13]. At the micro-evolutionary level, individuals are described in terms of behavioral traits (that could be socially accepted or unacceptable) which are passed from generation to generation using several socially motivated operators. At the macro-evolutionary level, individuals are able to generate "mappa", or generalized descriptions of their experiences. Individual mappa can be merged and modified to form "group mappa" using a set of generic or problem specific operators. Both levels share a communication link.

The micro-evolutionary level refers to the knowledge acquired by individuals through generations which are stored to guide the behavior of the individuals. This acquired knowledge is stored in the search space called belief space in CA during the evolution of the population. Interaction between the two basic components i.e., population space and Belief space make cultural algorithm as a dual inheritance system. Population space is that where the information about individuals is stored and the belief space is where the culture knowledge is formed and maintained during the evolution of the population.

The frame work of CA can be described as shown in Fig.1.

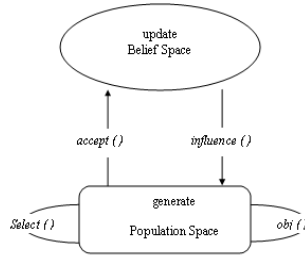


Fig. 1. Frame work of Cultural Algorithm

An acceptance function $accept()$ and updating function $update()$ play very vital role in belief space. After evolution of population space with a performance function $obj()$, $accept()$ will determine which individuals are kept aside for Belief space. Experiences of those elite individuals will update the knowledge of the Belief space via $update()$. These updated knowledge are used to influence the evolution of the population.

A pseudo- code description of cultural algorithm is described as follows,

Begin

$t = 0$

Initialize P^t

Initialize B^t

Repeat

Evaluate P^t

Update ($B^t, accept(P^t)$)

Generate ($P^t, influence(B^t)$)

$t = t + 1$

Select P^t from P^{t-1}

Until (Termination condition achieved)

end

4 Overview of the Proposed Approach

The basic idea of using CA with evolutionary programming is to influence the mutation operator so that the current knowledge stored in the search space can be properly exploited. Cultural Algorithm belongs to the class of evolutionary algorithms which offers a unique strategy for optimization. The strategy used here is described by the following steps.

4.1 Initialization

The optimization process in CA is carried with the basic operations: initialization, modification of belief space, mutation and selection. The algorithm starts by creating a population vector P of size N_p composed of individuals that evolve over G

generations. Each individual X_i is a vector that contains as many elements as the problem decision variable. The population size N_p is an algorithm control parameter selected by the user. Thus,

$$P^{(G)} = [X_i^{(G)}, \dots, X_{N_p}^{(G)}]$$

$$X_i^{(G)} = [X_{1,i}^{(G)}, \dots, X_{D,i}^{(G)}]$$

$$i = 1, \dots, N_p$$

The initial population is chosen randomly in order to cover the entire searching region uniformly. A uniform probability distribution for all random variables is assumed in the following as

$$X_{j,i}^{(0)} = X_j^{\min} + \sigma_j (X_j^{\max} - X_j^{\min})$$

Where $i = 1, \dots, N_p$ and $j = 1, \dots, D$

Here D is the number of decision or control variables, X_j^{\min} and X_j^{\max} are the lower and upper limits of the j th decision variable and $\sigma_j \in [0,1]$ is a uniformly distributed random number generated anew for each value of j . $X_{j,i}^{(0)}$ is the j th parameter of the i th individual of the initial population.

4.2 Modification of Belief Space

The acceptance function controls the information flow from the population space to the belief space. The acceptance function determines which individuals and their behavior impact the belief space knowledge. Top individuals are selected to update the belief space.

The parameter values for the current selected individuals by the acceptance function are used to calculate the current acceptable interval of normative knowledge. So, the update of normative knowledge is as follows. Assuming X_i and X_k be the individuals with minimum and maximum values for parameter j between the accepted individuals in the current generation, then

$$l_j^{t+1} = \begin{cases} X_{i,j}^t, & \text{if } X_{i,j} \leq l_j^t \text{ or } f(X_{i,j}^t) < L_j^t \\ l_j^t & \text{otherwise} \end{cases}$$

$$L_j^{t+1} = \begin{cases} f(X_{k,j}^t) & \text{if } X_{k,j} \leq l_j^t \text{ or } f(X_{k,j}^t) < L_j^t \\ L_j^t & \text{otherwise} \end{cases}$$

and,

$$u_j^{t+1} = \begin{cases} X_{k,j}^t, & \text{if } X_{k,j} \leq u_j^t \text{ or } f(X_{k,j}^t) < U_j^t \\ u_j^t & \text{otherwise} \end{cases}$$

$$U_j^{t+1} = \begin{cases} f(X_{k,j}^t) & \text{if } X_{k,j} \leq u_j^t \text{ or } f(X_{k,j}^t) < U_j^t \\ U_j^t & \text{otherwise} \end{cases}$$

Where, l_j^t represents lower bound for parameter j at generation t and L_j^t denotes the performance score for it and u_j^t represents upper bound for parameter j at generation t and U_j^t denotes the performance score for it.

4.3 Mutation Operation

Mutation takes place for each variable of each individual, with the influence of the belief space. If the variable j of the parent is outside the interval given by the normative part of the constraints, then we attempt to move within such interval through the use of a random variable. *The current individual of n numbers of candidate for parameter j* can be selected by the formula given,

$$X_{i+n,j} = \begin{cases} X_{n,j} + |(u_j - l_j) * N_{n,j}(0,1)| & \text{if } X_{n,j} < l_j \\ X_{n,j} - |(u_j - l_j) * N_{n,j}(0,1)| & \text{if } X_{n,j} < u_j \end{cases}$$

u_j and l_j represent the upper value and lower value of parameter j of current elite in the belief space.

4.4 Selection Operation

Selection is the operation through which better offspring are generated. To improve the speed of the algorithm, we take advantage of the rules for performing tournament selection. After performing mutation, we will have a population of size $2p$ (p parents generate p children). Tournament is performed considering the entire population. Tournaments consists of c confrontations per individual, with the c opponents randomly chosen from the entire population. When the tournaments finish, the p individuals with the largest number of victories are selected to form the following generation.

The optimization process is repeated for several generations. The iterative process of updating of belief space, mutation, and selection on the population will continue until a user-specified stopping criterion, normally, the maximum number of generations allowed, is met.

5 Structure of Solutions

In this section, an algorithm based on a cultural algorithm for optimal solution of economic load dispatch problem is described. For any population based algorithm the representation of individuals and their elements is very important. For the present problem, it is the candidate power generations of thermal units. The algorithm starts with the initialization process. Let $P^{(0)} = [X_1^{(0)}, X_2^{(0)}, \dots, X_k^{(0)}, \dots, X_{N_p}^{(0)}]$ be the initial population of N_p number of particles. For a system of n number of candidate generator, position of k th individual is of n -dimension and can be represented by $X_k^{(0)} = [PG_{k,1}^{(0)}, PG_{k,2}^{(0)}, \dots, PG_{k,j}^{(0)}, \dots, PG_{k,n}^{(0)}]$

The element $PG_{k,j}^{(0)}$ represents a randomly selected power generation satisfying the constraints given by (3).

6 Simulation Results

The proposed algorithm has been applied on a sample test system to verify its feasibility and effectiveness. The algorithm has been written in MATLAB and run a 3.0 MHZ, 1GB RAM PC. The test system consists of 13 thermal generating with the effects of valve point loading .Cost coefficients and generation limits of thirteen units System are taken from [4].

Table 1. Results for 13-Generators system

Unit	Generation (MW)	Unit	Generation (MW)
P ₁	408.4368	P ₈	58.9985
P ₂	222.6675	P ₉	111.5505
P ₃	262.9531	P ₁₀	73.6298
P ₄	80.9229	P ₁₁	69.8230
P ₅	139.9683	P ₁₂	70.0545
P ₆	100.7569	P ₁₃	62.1934
P ₇	138.0447	Total Generation (MW)	1800

The parameter values are selected by trial and error method. The following values are selected for optimal results. Population size is 10, mutation probability is taken as 0.75, and maximum iterative generation number is 500. The load demand is taken as 1800 MW.

The optimal results are shown in Table 1. It is seen from Table 1, that optimal fuel cost is obtained 17683(\$/h). Table 2 compares the solution obtained by the proposed method with other methods like Particle Swarm Optimization [17], Genetic Algorithm [2], Evolutionary Programming [15] etc. It is seen that the best result using CA is comparatively lower than the other studies presented here.

Fig.2 shows convergence characteristics for a demand of 1800 MW.

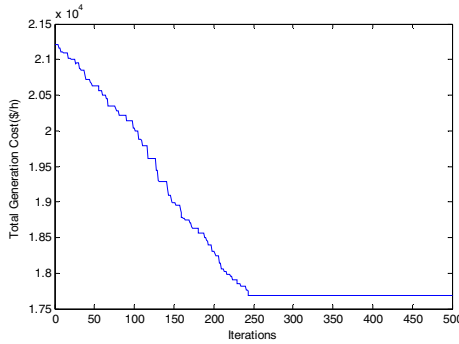


Fig. 2. Convergence characteristics of fuel cost

Table 2. Comparison of best fuel costs among different methods

Optimization Technique	Fuel Cost (\$/h)
PSO	18030.72
EP	17994.07
GA	17975.3437
DE	17963.83
CDEMD	17961.944
CA	17683.00

7 Conclusions

Economic Load Dispatch is one of the important problem for power systems operation. In this paper, a novel cultural algorithm has been proposed to solve Economic Load Dispatch problems. The proposed method has been tested on 13-generator systems with valve point effects. The results obtained by the proposed method are compared with other methods like GA, DE, EP, PSO etc. It is found that that proposed method can produce comparable results.

References

1. Chen, C.L., Wang, S.C.: Branch and bound scheduling for thermal generating units. *IEEE Trans. on Energy Conversion* 8(2), 184–189 (1993)
2. Chiang, C.L.: Improved Genetic Algorithm for power economic dispatch of units with valve-point effects and multiple fuels. *IEEE Transactions on Power Systems* 20(4), 1690–1699 (2005)
3. Coelho, L.D.S., Souza, R.C.T., Mariani, V.C.: Improved differential evolution approach based on cultural algorithm and diversity measure applied to solve economic load dispatch problems. *Mathematics and Computers in Simulation* 79, 3136–3147 (2009)
4. Chen, P.H., Chang, H.C.: Large-scale economic dispatch by genetic algorithm. *IEEE Trans. Power Syst.* 10(4), 1919–1926 (1995)
5. Chung, C.-J., Reynolds, R.G.: The Use of Cultural Algorithms Supprt Self-adaptove in EP. In: *Proc. of 1996 Adaptive Distributed Parallel Computing Symposium* (1996)

6. Chung, C.-J., Reynolds, R.G.: CAEP: An Evolution-based Tool for real valued Function Optimization using Cultural algorithms. *International Journal on Artificial Intelligence Tools* 7(3) (1998)
7. Jin, X., Reynolds, G.R.: Using Knowledge –Based Evolutionary Computation to Solve Nonlinear Constraint Optimization Problem: a Cultural Algorithm Approach. In: *Proc. of 1999 Congress on Evolutionary Computation*, pp. 1672–1678 (1999)
8. Gaing, Z.-L.: Particle swarm optimization to solving the economic dispatch considering generator constraints. *IEEE Trans. Power Syst.* 18(3), 1718–1727 (2003)
9. Giridhar, K., Mouly, V.S.R.K.: Using Evolutionary Computation to solve the Economic Load Dispatch problem. *IEEE Transactions on Power Systems* 3, 296–301 (2001)
10. Lee, K.Y., et al.: Fuel cost minimization for both real and reactive power dispatches. *IEE Proc. C, Gener. Trns. & Distr.* 131(3), 85–93 (1984)
11. Noman, N., Iba, H.: Differential Evolution for economic load dispatch problem. *Electric Power Systems Research* 78(3), 1322–1331 (2008)
12. Park, J.H., Yang, S.O., Mun, K.J., Lee, H.S., Jung, J.W.: An Application of Evolutionary Computations to Economic Load Dispatch with piecewise Quadratic Cost Functions. *IEEE Transactions on Power Systems* 69, 289–294 (1998)
13. Reynolds, R.G.: An Introduction to Cultural Algorithms. In: *Proceedings of the 3rd Annual Conference on Evolutionary Programming*, pp. 131–139. World Scientific Publishing (1994)
14. Reynolds, R.G.: An Overview of Cultural Algorithms. In: *Advances in Evolutionary Computation*. McGraw Hill Press (1999)
15. Sinha, N., Chakrabarti, R., Chattopadhyay, P.K.: Evolutionary Programming Techniques for economic load dispatch. *IEEE Trans. on Evolutionary Computation* 7(1), 83–94 (2003)
16. Song, Y.H., Wang, G.S., Wang, P.Y., Johns, A.T.: Environmental/Economic Dispatch using Fuzzy Logic Controlled Genetic Algorithm. *Proc. I.E.E Generation, Transmission and Distribution* 144(4), 377–382 (1997)
17. Victoire, T.A.A., Jeyakumar, A.E.: Hybrid PSO-SQP for economic dispatch with valve-point effect. *Electric Power System Research* 71(1), 51–59 (2004)
18. Wood, A.J., Wollenberg, B.F.: *Power Generation, Operation and Control*. Wiley, New York (1984)
19. Yang, H.T., Yang, P.C., Huang, C.L.: Evolutionary programming based economic dispatch for units with nonsmooth fuel cost functions. *IEEE Trans. Power Syst.* 11(1), 112–118 (1996)

New Measure of Interestingness for Efficient Extraction of Association Rules

Parvati Bhurani, Mushtaq Ahmed, and Yogesh Kumar Meena

Malaviya National Institute of Technology, Jaipur, India
parvati_bhurani@rediffmail.com, mushtaqahmed@mnit.ac.in,
yogimnit@gmail.com

Abstract. Data Mining helps to uncover the already unknown and non-redundant knowledge in large databases, which can be used for decision making purpose. Association rule mining is one of the key research area in the field of Data Mining. Association rule mining can be considered as unsupervised learning model, it discovers the interesting relationship among large set of data items on the basis of some predefined threshold. Support-confidence is the classical model used for the rule mining purpose, it uses confidence for final rule generation but it has some limitations. As sometimes it can generate those rules which are not positively correlated and thus can mislead the decision maker. In this paper we addressed the problems associated with existing approach and also proposed two new measure of interestingness to deal with these problems. The new measures have been tested for their correctness.

Keywords: Association Rules, Interestingness Measure, Support-confidence, Correlation.

1 Introduction

Data Mining plays an important role to extract out of sight or concealed patterns from large datasets. The aim of Data Mining technique is to uncover the previously unknown, useful, and non-redundant cognition in large databases. It has a potential to help companies to focus on the most important information in their data warehouse. It has been used in many industries like retail market, insurance and banking etc. to increase sales, risk assessment and many more. Data Mining is also referred as knowledge discovery in databases (KDD).

During mining large number of rules are generated but only the small set has significance for the user point of view, so to find whether a rule is of interest or not it must need a suitable metric to measure the degree of rule in which the user is interested. Thus the role of interestingness measure plays an important role in rule extraction process [3]. Association rule mining or frequent pattern mining is a customary and well explored method for discovering interesting relations between itemsets in large data repository under the domain of Data Mining. The concept was first introduced by Agrawal [1,2] to analyze the data of a supermarket containing large collection of customer transaction. It is

a kind of unsupervised learning technique. The following is the general form of an association rule:

$$X \Rightarrow Y$$

In the above rule X is known as antecedent and Y is known as consequent and it can be read as if X then Y. It measures the association between X and Y. Support-confidence model [2] for association rule can be described as given: Let $I = i_1, i_2, \dots, i_n$ be a set of n literals called items and T be a set of transactions, where each transaction $t \in T$ is a set of items such that $t \subseteq I$. Each transaction is associated with a unique id TID. An association rule is the implication of the form, $X \Rightarrow Y$, where $X \subseteq I, Y \subseteq I$ and $X \cap Y = \emptyset$. Here X is the antecedent and Y is the consequent. Terms used in rule mining process defined are as follows:

Support(sup): Support of an itemset can be defined as the ratio of transactions containing the items in both antecedent and consequent of the rule to the total number of transaction. It measures the strength of the given itemset. Let $X \Rightarrow Y$ be the rule then support is given as below:

$$sup(X \Rightarrow Y) = \frac{sup_count(X \cap Y)}{T} \quad (1)$$

Here sup_count is the number of transactions containing the given itemset $X \cap Y$.

Confidence(conf): Confidence of an association rule measures how often items in Y appear in transaction that contains X. It measures the strength of a given association rule.

$$conf(X \Rightarrow Y) = \frac{sup(X \cup Y)}{sup(X)} \quad (2)$$

The rule mining process works in two steps:

Frequent itemset identification: The itemset which satisfy the minimum support threshold (σ) are generated in this phase.

Rule extraction: This phase takes frequent itemset as input which has been generated in step 1 along with the value of minimum confidence threshold (τ). The rules which satisfy the minimum confidence are extracted in this phase.

The remaining paper is organized as follows: In section 2, some of the related work in this area is presented. Limitation in existing approaches are discussed in section 3. In section 4 we present the new measures of interestingness while in section 5, new measure are tested on sample dataset. Conclusion is given in last section.

2 Related Work

Association rule mining is one of the actively researched area in the Data Mining community. Lot of work has been reported in this domain and used in various applications in real world. The algorithms discussed here are generally variation of classical apriori approach and involves a time consuming candidate generation process. Apriori [1] is one of important algorithm for mining the frequent

itemset. The process of rule discovery is divided into two subproblems the first is to find the frequent itemset with minimum support threshold while the second is to find the association rules from the frequent itemset generated in first phase with the help of minimum confidence. Some variations of apriori are also proposed like partition based [6], sampling technique [7] to deal with the associated problems. Two major issues are observed in above methods as all the addressed approaches are variation of classical apriori algorithm thus involves the time consuming candidate generation process and these approaches also suffer from a rare item problem because of single minimum support threshold. These two issues are discussed by many researchers and solutions are also proposed to deal with these issues. FP-Growth [9], it uses a new tree data structure for storing compressed information regarding the frequent patterns. FP-growth is better than other approaches as FP-tree is constructed which removes the requirement of costly database scan and a pattern growth approach avoids costly candidate generation and thus it is time efficient. CFP-Growth, it is the extension of FP-growth algorithm and uses CFP-tree structure in place of FP-tree for storing the frequent itemset. The method discussed in this section works as per traditional support-confidence model and thus suffers from limitation like generation of rules which are not important as per the user criteria. We discuss this problem in section 3 with some example.

3 Limitation of Support-Confidence Framework

Support-confidence framework uses confidence measure for the rule generation process, as in some cases it may discover uncorrelated rules which can mislead the decision maker. Let us consider the $\sigma = 20\%$ and $\tau = 40\%$. Following is the sample dataset taken from supermarket illustrates the above addressed limitation: From the above data we can have following set of association rules:

Table 1. Sample Dataset

<i>Item/Tran_No</i>	1	2	3	4	5	6	7	8	9	10
Bread	1	1	1	1	0	0	0	0	0	0
Jam	1	1	1	0	1	1	1	1	1	0
Egg	0	1	1	0	1	1	1	0	0	0
Butter	1	0	0	1	0	1	1	1	1	1

- Rule1: bread \Rightarrow jam [30%, 75%]
- Rule2: bread \Rightarrow egg [20%, 50%]
- Rule3: jam \Rightarrow egg [50%, 62.5%]
- Rule4: jam \Rightarrow butter [50%, 62.5%]

In the above mentioned rules, first parameter is the support of the rule while the second parameter is the confidence of the rule. As per the confidence value the

first rule is very strong but in reality there is negative correlation between bread and jam as support of jam alone is 80% which is greater than the confidence value, while the third rule has lower confidence than the first rule but it has a positive correlation between jam and egg as the support of egg alone is 50% which is lower than the confidence value and thus if we associate egg with jam then support of egg will be increased. Now consider second rule then we can observe that there is no association between the two variables i.e. bread and egg are independent to each other as support of egg is 50% which is equal to the confidence value.

Suggested Solution: The above example suggests that it is not necessary that all strong rules i.e. the rules with high confidence value are interesting. The problem occurs because in confidence we do not consider the baseline frequency of consequent. So to overcome this problem base line frequency of consequent is to be taken into consideration while measuring the correlation among different itemset. In following section we will see the new interest measure with their properties.

4 Proposed Objective Interest Measure

In this section we introduce two new measure of interestingness namely ratioPS and ratioLEV. We discuss these two new measures along with their set of properties and possible range. Let us consider $X \Rightarrow Y$ as an association rule.

ratioPS: It is an asymmetric interest measure and hence the value of ratioPS(XY) is different from ratioPS(YX). Following is the formula for this new measure:

$$ratioPS = \frac{P(XY) - P(X) * P(Y)}{1 - P(X)} \tag{3}$$

The possible range of this measure is from -1 to +1. We can consider this measure as the ratio of PS and the probability of \bar{X} .

Table 2. Correlation criteria between two itemset X and Y

S.No.	Value	Correlation
1	If ratioPS($X \Rightarrow Y \prec 0$)	Negatively Correlated
2	If ratioPS($X \Rightarrow Y = 0$)	No Correlation
3	If ratioPS($X \Rightarrow Y \succ 0$)	Positively Correlated

ratioLEV: This measure can be considered as the ratio of leverage to the product of probability of \bar{X} and Y.

$$ratioLEV = \frac{\frac{P(XY)}{P(X)} - P(X) * P(Y)}{(1 - P(X)) * P(Y)} \tag{4}$$

The possible range of this measure can vary from 0 to ∞ . Let us consider a transaction dataset of 10 transactions, this example illustrate how the new measure identify the true associations between two itemsets namely X and Y:

Table 3. Correlation between two itemset X and Y

S.No.	Value	Correlation
1	If $\text{ratioLEV}(X \Rightarrow Y) < 1$	Negatively Correlated
2	If $\text{ratioLEV}(X \Rightarrow Y) = 1$	No Correlation
3	If $\text{ratioLEV}(X \Rightarrow Y) > 1$	Positively Correlated

Table 4. Results of the new interest measure

S.No.	P(XY)	P(X)	P(Y)	Conf.	ratioPS	ratioLEV	Remark
1	0.3	0.4	0.8	0.75	-0.33	0.89	Negative Correlated
2	0.2	0.4	0.5	0.5	0.0	1.0	Independent
3	0.4	0.4	0.7	1.0	1.7	0.2	Positive Correlated

From the above example we can see that both the proposed measures are capable to detect the negatively correlated itemset.

4.1 Properties Followed by the New Measures

In this section we see the set of properties which should be followed by various interest measures and then we will show the properties satisfied by new measures. Following is the description of properties:

Property 1: This property checks whether the two itemset are independent or not. If both X and Y are independent then the value of $P=0$.

Property 2: If value of joint probability i.e. $P(X,Y)$ increases when the antecedent ($P(X)$) and consequent ($P(Y)$) probability remain unchanged. Then the value of P increases.

Property 3: If $P(X)$ decreases when $P(Y)$ and $P(X,Y)$ remain unchanged or $P(Y)$ decreases when $P(X)$ and $P(X,Y)$ remain unchanged. Then the value of P increases.

The above properties proposed by Piatetsky-Shapiro [4] the first property can be relaxed as some measure has $P=1$ when the two itemset are independent.

Property 4: The property is known as symmetry under variable permutation, If the itemset in the antecedent and consequent are exchanged then there is no change in value of P before and after the permutation.

Property 5: The property is known as scaling invariance under row or column, If row or column values in the contingency table are multiplied by some factor then no change in value of P after row or column scaling.

Property 6: The property is known as antisymmetry under row or column permutation, If the row or column values in the contingency table are swapped then P becomes -P.

Property 7: The property is known as invariance under inversion, If both row and column values in the contingency table are swapped at the same time then no change in value of P.

Property 8: The property is known as invariance under the addition of null record, In this new transactions are added which do not contain any of the itemset under consideration. No change in value of P after null addition. The above properties(4-8) proposed by Tan et al. [5] which are based upon operations for 2x2 contingency matrix.

Next we present the set of properties followed by the measure ratioPS and ratioLEV in the following table:

Table 5. Properties followed by proposed measures

Measure/Property	P1	P2	P3	P4	P5	P6	P7	P8
ratioPS	Y	Y	Y	Y	Y	N	N	N
ratioLEV	N	Y	Y	Y	Y	N	N	N

Here Y means the property is satisfied while N represents that the property is not satisfied by the particular measure.

5 Testing of New Measures on Sample Dataset

Sample Dataset: Here, we present the interest value of the new measures by using pearson’s correlation coefficient measure along with some of the existing measures on sample dataset. Let us consider the following dataset:

Table 6. Sample Dataset For Illustration of Proposed Approach

Tran No	Itemset
1	Beef, Chicken, Milk
2	Beef, Cheese
3	Cheese, Boots
4	Beef, Chicken, Cheese
5	Beef, Chicken, Milk, Cheese, Clothes
6	Chicken, Milk, Clothes
7	Chicken, Milk, Clothes
8	Beef, Chicken

Following is the formula for pearson correlation coefficient:

$$r = \frac{\sum XY - (\sum X)(\sum Y)}{\sqrt{(\sum X^2 - \frac{\sum X^2}{N})(\sum Y^2 - \frac{\sum Y^2}{N})}} \quad (5)$$

Here r can take three values and on the basis of these values we can identify whether the two itemset (X and Y) are independent, positively correlated or negatively correlated. If r=0 means both X and Y are independent, if r is greater than 0 means X and Y are positively correlated otherwise X and Y are negatively correlated.

Table 7. Interest values by different measures

Association Rule	Conf.	Lift	PS	CR	AV	ratioPS	ratioLEV
Beef \Rightarrow Chicken	.66	.88	-.62	-.33	-.083	-.25	0.55
Chicken \Rightarrow Beef	.66	.88	-.62	-.33	-.083	-.25	0.55
Beef \Rightarrow Cheese	0.5	1.0	0.0	0.0	0.0	0.0	1.0
Cheese \Rightarrow Beef	0.75	1.0	0.0	0.0	0.0	0.0	1.0
Chicken \Rightarrow Milk	0.66	1.33	0.125	0.57	0.16	0.5	2.33
Milk \Rightarrow Chicken	1.0	1.33	0.125	0.57	0.16	0.5	1.66
Clothes \Rightarrow Chicken	1.0	1.33	0.093	0.447	0.25	0.15	1.53
Milk \Rightarrow Clothes	0.75	2.0	0.18	0.77	0.375	.375	3.0
Clothes \Rightarrow Milk	1.0	2.0	0.18	0.77	0.5	0.3	2.6
{Beef,Milk} \Rightarrow Chicken	0.5	0.66	-.125	-0.57	-.25	-.25	0.33
{Beef,Cheese} \Rightarrow Chicken	0.5	0.66	-.125	-0.57	-.25	-.25	0.33
{Chicken,Cheese} \Rightarrow Beef	0.5	0.66	-.125	-0.57	-.25	-.25	0.33
Milk \Rightarrow {Chicken,Clothes}	0.75	2.0	0.18	0.77	0.375	0.375	3.0
Clothes \Rightarrow {Chicken,Milk}	1.0	2.0	0.18	0.77	0.5	0.3	2.6
{Chicken,Milk} \Rightarrow Clothes	0.75	2.0	0.18	0.77	0.375	0.375	3.0
{Chicken,Clothes} \Rightarrow Milk	1.0	2.0	0.18	0.77	0.5	0.3	2.6
{Milk,Clothes} \Rightarrow Chicken	1.0	1.33	0.093	0.447	0.25	0.15	1.53

From the table 7, it can be seen that the proposed measures are equivalent to some of the existing measures, like lift, PS, correlation, which also taken consideration of the negative correlation. As per the given dataset we can see that some rules are negatively correlated like, *beef* \Rightarrow *chicken*, *chicken* \Rightarrow *beef* and *beef*, *chicken* \Rightarrow *cheese* etc., all these negatively correlated rules are not detected by the classical support-confidence measure and thus we infer that the proposed measure are capable to detect all kinds of correlation hence can help in correct decision making.

6 Conclusion

The work proposed in this paper introduces a new measure of interest to extract the association rules. It deals with the limitation of existing support-confidence

framework. In this paper we have introduced two measure of interestingness and have seen that both are capable to capture negative correlation among different itemsets. We have seen some properties followed by the proposed measure and correctness of new measures also checked on sample dataset.

References

1. Imielinski, T., Agrawal, R., Swami, A.N.: Mining association rules between sets of items in large databases. In: Proceedings of the ACM SIGMOD Conference on Management of Data, vol. 22, pp. 207–216 (1993)
2. Agrawal, R., Srikant, R.: Fast algorithm for mining association rules. In: Proceeding of 20th International Conference on Very Large Databases, pp. 487–499 (2003)
3. Silberschatz, A., Tuzhilin, A.: What makes pattern interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering*, 970–974 (1996)
4. Piatetsky-Shapiro, G.: Discovery, analysis and presentation of strong rules. In: Piatetsky-Shapiro, G., Frawley, W. (eds.) *Knowledge Discovery in Databases*, pp. 229–248 (1991)
5. Kumar, V., Tan, P., Srivastva, J.: Selecting the right interestingness measure for association patterns. In: Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining, pp. 32–41 (2002)
6. Omiecinski, E., Savasere, A., Navathe, S.: An efficient algorithm for mining association in large databases. In: Proceedings of the 21st International Conference on Very Large Databases, pp. 432–444 (1995)
7. Toivonen, H.: Sampling large databases for association rules. *VLDB Journal*, 134–145 (1996)
8. Ullman, J.D., Brin, S., Motwani, R., Tsur, S.: Dynamic itemset counting and implication rules for market basket data. In: Proceedings of ACM SIGMOD International Conference Management of Data, vol. 8, pp. 255–264 (1997)
9. Pei, J., Han, J., Yin, Y.: Mining frequent patterns without candidate generation. In: Proc. ACM-SIGMOD International Conference on Management of Data, pp. 1–12 (2000)
10. Zhang, C., Wu, X., Zhang, S.: Efficient mining of both positive and negative association rules. *ACM Transaction on Information Systems* 22, 381–405 (2004)
11. Geng, L., Hamilton, H.J.: Interestingness measures for Data Mining. A survey. *ACM Computing Surveys* 38 (2006)
12. Vanhoof, K., Brijs, T., Vets, G.: Defining interestingness for association rules. *International Journal on Information Theories Applications* 10, 370–375 (2010)

Improving Prediction of Interdomain Linkers in Protein Sequences Using a Consensus Approach

Piyali Chatterjee^{1,*}, Subhadip Basu², Mahantapas Kundu², and Mita Nasipuri²

¹ Department of Computer Science and Engineering, Netaji Subhash Engineering College, Garia, Kolkata – 700152, India

² Department of Computer Science and Engineering, Jadavpur University, Kolkata – 700032, India

Abstract. A two pronged strategy, one involving consensus approach and the Multi Layer Perceptron (MLP) as the classifier and the other including physico-chemical properties as additional features, is proposed and implemented here for improved prediction of interdomain linkers in protein chains. The software is tested on proteins of the CASP6 experiment in order to measure its prediction accuracy using three-fold cross validation. Finally, our consensus approach combines results of 28 different neural networks. We observe significant improvements of AUC scores by 9.4% on average in comparison to the corresponding most successful single artificial neural networks.

1 Introduction

A *domain* is a segment of a polypeptide chain that can fold into a three dimensional structure irrespective of the presence of other segments of the chain [1]. The overall 3D structure of the polypeptide chain is referred to as the protein's tertiary structure and the domain is the fundamental building block of such tertiary structure. To predict the tertiary structure of a protein, it is useful to segment the protein by identifying domain boundaries in it. The knowledge of domains is used to classify proteins and understand their structures, functions and evolution.

A number of methods so far have been developed to identify protein domains starting from their primary sequences. All these are mainly developed for prediction of multi-domains in protein chains. Out of the protein domain prediction methods, developed in the last five years or so, DOMpro [2] is an important one. It employs the machine learning algorithms in the form of recursive neural networks (1D- RNNs) to predict domains in a protein chain. Moreover, a combination of evolutionary information in the form of profiles, predicted secondary structures, predicted solvent accessibility of the protein chains are utilized. DOMpro is experimentally observed to have correctly predicted the domains from the combined dataset of single and multi-domain proteins in 69% of the cases. Armadillo [3], another domain predictor, uses *Domain Linker propensity Index* (DLI) to convert a protein sequence to a smoothed numeric profile, from which domains and domain boundaries may be predicted. The work is finally reported to have achieved 37% sensitivity for multi-domain proteins.

* Corresponding author.

The position specific scoring matrix of the target protein, which is obtained through PSI-BLAST, has also been used for domain boundary prediction by PPRODO [4]. A neural network has been finally used there as a classifier. The overall accuracy of domain boundary prediction as achieved by PPRODO [4] is 67%. CHOPnet [5] uses evolutionary information, predicted one-dimensional structure (secondary structure, solvent accessibility), amino acid flexibility and amino acid composition for predicting domains in protein chains. It finally predicts domains in a protein chain by removing noisy peaks from the raw outputs of the neural networks, used for this work, through a process of post processing. A prediction accuracy of 69% on all proteins is reported for this work. Galzitskaya *et al* [6] [7] have developed a method based on finding the minima in a latent entropy profile. This method correctly predicts the domain boundaries for about 60% proteins [7]. In the work of Sikder and Zomaya [8], the performance of DomainDiscovery of protein domain boundary assignment is improved significantly by including the inter domain linker index value along with an inter domain linker index values, Position Specific Scoring Matrix (PSSM), predicted secondary structures, solvent accessibility information. Support Vector Machine (SVM) is used to predict possible domain boundaries for target sequences. The method is reported to have achieved 70% accuracy for multi-domain proteins. Based on the difference in amino acid compositions between domain and linker regions, a method DOMCUT [9] has been developed to predict linker regions among domains. The sensitivity and the selectivity, as achieved by this method, are 53.5% and 50.1% respectively.

Based on the application of secondary structure element alignment (SSEA) and profile-profile alignment (PPA) in combination with InterPro pattern searches, a protein domain prediction approach, called SSEP-Domain, is proposed by Gewehr and Zimmer [10]. A preliminary version of SSEP-Domain is ranked among the top-performing domain prediction servers in the CASP6 and CAFASP4 experiments. Cheng [11] proposed a hybrid domain prediction web service, called DOMAC, by integrating *template-based* and *ab-initio* methods. The preliminary version of the server (DOMAC) is ranked among the top domain prediction servers in the CASP7, 2006. However, this performance is very likely to be overestimated. So, DOMAC is also evaluated on a larger, balanced, the high quality data set manually curated by Holland *et al.*[12]. As a result, the overall domain number prediction accuracies of the *template-based* and *ab-initio* methods are 75% and 46% respectively. To achieve a more accurate and stable predictive performance than the existing state-of-the-art models, a new machine learning based domain predictor, viz., DomNet [13] is trained using a novel compact domain profile, predicted secondary structure, solvent accessibility information and inter-domain linker index. Its performance is tested on the benchmark_2 dataset. It is observed to have achieved 71% accuracy.

From the above discussion on the past attempts for protein domain prediction it can be summarized best possible prediction accuracy ranges from 67% to 70% for the *ab-initio* methods discussed here.

In the light of the above discussion, it appears that there is still some scope for improvement in protein domain prediction. Physicochemical properties of amino acids in addition to usual features and consensus of artificial neural network based MLP classifiers appear to have potentials for implementation of this method. The rationale

behind the choices of the feature sets and classifiers for prediction of domain boundaries are discussed in the following sections.

2 Methods

An attempt has been made under the present work to employ Multi-layer Perceptron (MLPs) with varying hidden layer neurons, for protein domain prediction on the basis of powerful feature set. In this work, an MLP, a *feed-forward* model of the artificial Neural Network with *learning* and *generalization* abilities, is employed for providing prediction decisions on each of the two classes, viz., linker, *non-linker* or *domain* of each residue of an amino acid sequence. In doing so, a 29-residue window is slid over the protein chain every time by one residue position until the end of chain is covered by the window. At a particular portion of the window, all the above-mentioned features are extracted for each residue covered by the window. The feature values are then fed into the input of the MLP, which decides about whether the central residue of the window is part of a domain or a linker (i.e., a non-domain) by making its output positive or negative (1 or 0). The process is repeated until the window covers the last residue of the protein chain under consideration and in this process, domain and linker regions in the protein chain are identified

2.1 Biophysical Descriptors

Five types of biophysical descriptors, i.e. features describing each of existing amino acids, namely, predicted ordered or disordered region, Normalized flexibility parameters (B-values), polarity, linker index, Modified Kyte-Doolittle hydrophobicity scale are used for this work.

2.1.1 Predicted Ordered or Disordered Region

Disordered region is found to be a good approximation of global structural analysis of the domain arrangement in the three dimensional structure of multi-domain proteins. From experimental findings, it is known that large ordered region when they are divided by shorter parts of disordered regions in a protein chain, are likely to be separate domains [14]. Disordered local sequence segments are likely to be linkers or inter domain spacers between protein domains. In this work, Disport (VL3H, window size 11) tool has been used to identify ordered or disordered region. This ordered or disordered region has been used as one of the features.

2.1.2 Normalized Flexibility Parameters (B-Values)

The presence of multiple domains in proteins gives rise to a great deal of flexibility and mobility, leading to protein domain dynamics. Flexible regions are considered to be natively unfolded. The Debye-Waller factor (B-value), which measures local residue flexibility, is widely used to measure residue flexibility. The prediction of flexibility may help to unravel protein function. Here, Normalized average flexibility parameters (B-values) from the AAINDEX dataset [15] have been taken as another feature as presence of multiple domains increases protein flexibility.

2.1.3 Polarity

The distribution of polar and non-polar side chains is the most important factor governing the folding of a protein into 3D structure. Since domain is a unit of 3D structure, polarity has been taken as a feature from the AAINDEX dataset [15].

2.1.4 Amino Acid Linker Index

To represent the preference for amino acid residues in linker or regions, a parameter called the *linker index* is defined by Sumaya and Ohara [9]. This we have used here as a feature. From the AAINDEX dataset, linker index has been used as an important feature.

2.1.5 Hydrophobicity

Folding is driven by the burial of hydrophobic side chains into the interior of the molecule so to avoid contact with aqueous environment. The average hydrophobicity for linkers is 0.65 ± 0.09 . Small linkers show an average hydrophobicity of 0.69 ± 0.11 , while large linkers are more hydrophobic with 0.62 ± 0.08 . The more exposed the linker, the more likely it is to contain hydrophilic residues. Greater hydrophobicity is found in more linker connections between two domains. Modified Kyte-Doolittle hydrophobicity scale is taken as a feature, which is also from the AAINDEX dataset.

3 Results

The current experiment is conducted in two stages. In the first stage 354 protein chains of the CATH database (version 2.5.1) are used to perform a three-fold cross validation experiment. MLP based classifiers, designed to produce optimum Area under Receiver Operating Characteristics (ROC) Curve (AUC). The evaluation measures, considered here for testing performances of the present technique, are Area under ROC curve, Recall and Precision. We thus generate a pool of trained networks from the hidden neuron variations in three cross-validated experiments. In this work, network weights corresponding to a particular hidden layer neuron are saved while optimizing AUC values for any given training and test dataset. More specifically, while optimizing AUC, the corresponding recall and precision scores are recorded. In the second stage of the experiment, we consider a consensus approach on the basis of the trained networks to generate test results on 19 protein sequences, taken from the CASP6 dataset. The following subsections discuss the detailed experimental protocol and the results obtained from these two stages of experiment.

3.1 Three Fold Cross Validation Experiment with CATH Database

The protein domain prediction technique presented here is applied on a curated data set, derived from 354 protein chains from the CATH database, version 2.5.1. Three fold cross-validation experiments are done on 354 protein chains of CATH database. In designing the MLP classifiers here, networks are trained several times by varying hidden layer neurons from 70 to 160 by setting 0.8 as learning rate and 0.7 as momentum term. Experiments are conducted for each such fold, by optimizing the training networks for best AUC. The detailed variation of network performances of the three cross-validation folds are shown in Table 1.

Table 1. Performance Measures on test sets of three cross-validation sets Optimizing AUC score

Cross validation Set	Topology	AUC (%)	Precision (%)	Recall (%)	Error (%)
CV1	145-150-2	65.79	47.28	40.65	17.44
CV2	145-160-2	61.25	41.04	31.06	18.98
CV3	145-150-2	64.55	38.96	42.39	15.06
Average	-	63.86	42.42	38.03	17.02

3.2 Consensus Prediction over CASP6 Dataset

To independently evaluate the strength of our technique 19 from CASP6 dataset are tested by our method respectively. AUC optimized training networks from three cross-validated folds are considered here. To further improve the prediction accuracy of single networks, a consensus approach has been considered here.

Table 2. AUC scores from 3 star consensus and single network for CASP6 target proteins

Targets	Maximum AUC obtained from single network	Maximum AUC obtained from 3star consensus	Gain (%)
T0226	0.633	0.9	26.7
T0272	0.643	0.8	15.7
T0280	0.703	0.857	15.4
T0279	0.857	1	14.3
T0199	0.847	0.984	13.7
T0233	0.727	0.837	11
T0249	0.78	0.889	10.9
T0216	0.567	0.667	10
T0209	0.684	0.769	8.5
T0228	0.675	0.76	8.5
T0262	0.707	0.753	4.6
T0237	0.525	0.544	1.9
T0202	1	1	0
T0222	0.998	0.998	0
T0247	1	1	0
T0248	0.971	0.971	0
T0268	0.5	0.5	0
T0235	0.521	0.519	-0.2
T0229	0.833	0.804	-2.9

In this approach, three best AUC optimized networks from the three cross-validation folds are considered for consensus. In each such fold, 10 variations in hidden neurons produce 10 trained networks. However, in the second cross validation fold for two hidden neuron variation (80 and 100), stable trained networks could not be obtained. Therefore we get 28 training networks for n-star quality consensus strategy. The results for 3-star consensus are listed in Table-2. It may be observed that in most cases the consensus approach improves the single best classification decision. The average and maximum gains of 3-star consensus over single network performances over CASP6 dataset are 7.26% and 26.7% respectively.

3.3 Comparison between Proposed Method and Other CAFASP-Dp Methods Consensus Prediction over CASP6 Dataset

In CAFASP4 experiment, there are 58 target proteins where 41 targets are single domain proteins and 17 targets are two-domain proteins. We took 19 targets from CASP6 experiments. 12 targets are common in CAFASP-DP experiments. Our prediction accuracies are comparable with those methods. For example, for target T0199, we achieved 97%, but Biozon achieved 98.92%. For targets T0202, T0249, T0262, T0272 we achieved better prediction accuracies than existing methods but for other targets, we achieved comparable accuracy than Biozon, Control1, Control2, Dompred-domssea, Robetta-rosettado, SSEP-DOMAIN methods. Some of them are presented in Table 3.

Table 3. Prediction results of other existing methods and proposed method for CAFASP-DP and CASP6 experiments

Targets	Armadillo	Biozon	Dompred-domssea	Dompro	SSEP-DOMAIN	Con-sensus	3* consensus	best single
T0199	X	98.82	91.42	74.56	88.17	89.94	97	84.7
T0202	X	61.04	61.85	61.85	61.85	61.85	100	100
T0209	60.25	75.73	50.63	85.77	88.7	89.12	91	68.4
T0216	66.21	54.02	51.49	75.63	73.33	99.08	96	56.7
T0222	59.25	34.32	98.39	84.72	96.78	93.3	97	100
T0228	50.58	54.55	95.8	55.94	61.77	77.62	94	67.5
T0229	92.03	68.12	68.12	68.12	68.12	68.12	88	83.3
T0233	20.99	59.67	89.78	97.51	90.06	94.48	97	72.7
T0235	59.92	65.33	75.95	53.51	96.59	47.9	86	52.1
T0249	63.64	68.42	93.3	61.72	91.39	99.52	97	78
T0262	67.58	68.36	67.97	89.45	60.16	83.98	94	70.7
T0272	82.46	98.58	59.24	59.24	59.24	59.24	97	64.3

4 Conclusions

We conclude that the designed feature set, alongside with MLP classifier based consensus approach effectively predicts the linkers and domain regions in multi-domain protein chains. Prediction decisions of AUC optimized networks from the three experimental folds are combined to design ‘3-star’ quality consensus strategy. In the consensus approach, 3-star quality consensus is designed by combining the decisions of the three best networks from each of the three sets of cross validation experiments. The consensus strategy is found to be superior in comparison with the performances of the best single networks.

Acknowledgments. Authors are thankful to the “CMATER”, of Computer Science & Engineering Department, Jadavpur University, for providing infrastructure facilities during progress of the work. One of the authors, Mrs. Piyali Chatterjee, is thankful to NSEC, Garia for kindly permitting her to carry on the research work.

References

1. Mount, D.: *Bioinformatics: Sequence and Genome Analysis*, 2nd edn., p. 416. Cold Spring Harbor Laboratory Press
2. Cheng, J., Sweredski, M.J., Baldi, P.: DOMpro: Protein Domain Prediction Using Profiles, Secondary Structure, Relative Solvent Accessibility, and Recursive Neural Networks. *Data Mining and Knowledge Discovery* 13, 1–10 (2006)
3. Dumontier, M., Yao, R., Feldman, H.J., Hogue, C.W.: Armadillo: Domain Boundary Prediction by Amino Acid Composition. *J. Mol. Biol.* 350, 1061–1073 (2005)
4. Sim, J., Kim, S., Lee, J.: PPRODO: Prediction of Domain Boundaries using Neural Networks. *Proteins* 59, 627–632 (2005)
5. Liu, J., Rost, B.: Sequence-based prediction of protein domains. *Nucleic Acid Research* 32, 3522–3530 (2004)
6. Galzitskaya, O.V., Melnik, B.S.: Prediction of protein domain boundaries from sequence alone. *Protein Science* 12, 696–701 (2003)
7. Galzitskaya, O.V., Dovidchenko, N.V., Lobanov, M.Y., Garbuzynskiy, S.O.: Prediction of Protein Domain Boundaries from statistics of Appearance of Amino Acid Residues. *Molecular Biology* 40(1), 96–107 (2006)
8. Sikder, A.R., Zomaya, A.Y.: Improving the performance of DomainDiscovery of protein domain boundary assignment using inter-domain linker index. In: *Proceedings of BMC Bioinformatics* (2006), doi:10.1186/1471-2105-7-S5-S6
9. Suyama, M., Ohara, O.: Domcut: prediction of inter-domain linker regions in amino acid sequences. *Bioinformatics* 19, 673–674 (2003)
10. Gewhr, J.E., Zimmer, R.: SSEP-Domain: protein domain prediction by alignment of secondary structure elements and profiles. *Bioinformatics* 22(2), 181–187 (2006)
11. Cheng, J.: DOMAC: an accurate, hybrid protein domain prediction server. *Nucleic Acids Research* 35, 354–356 (2007)
12. Holland, T., Veretnik, S., Shindyalov, I.N., Bourne, P.E.: A benchmark for domain assignment from protein 3-dimensional structure and its applications. *J. Mol. Biol.* 361, 562–590 (2006)

13. Yoo, P.D., Sikdar, A.R., Bing, Z., Taheri, B., Zomaya, A.Y.: DomNet: protein domain boundary prediction using enhanced general regression network and new profiles. *IEEE Transaction on NanoBioScience* 7(2), 172–181 (2008)
14. Wyrwicz, L.S., Koczyk, G., Rychleswski, L., Plewczynski, D.: ProteinSplit: splitting of multi-domain proteins using prediction of ordered and disordered regions in protein sequences for virtual structural genomics. *J. Physics Condensed Matter* 19 (2006)
15. Kawashima, S., Kanehisa, M.: AAINDEX: amino acid index database. *Nucleic Acids Res.* 28, 374 (2000)

Unconstrained Handwritten Digit OCR Using Projection Profile and Neural Network Approach

Amit Choudhary¹, Rahul Rishi², and Savita Ahlawat³

¹ Maharaja Surajmal Institute, New Delhi, India
amit.choudhary69@gmail.com

² Maharshi Dayanand University, Rohatk, Haryana, India
rahulrishi@rediffmail.com

³ Maharaja Surajmal Institute of Technology, New Delhi, India
savita.ahlawat@gmail.com

Abstract. The recognition accuracy of an Optical Character Recognition (OCR) system mainly depends on the selection of feature extraction technique and the classification algorithm. This paper focuses on the recognition of handwritten digits using projection profile features. Vertical, Horizontal, Left Diagonal and Right Diagonal directions are the four different orientations that are used for abstracting features from each handwritten digit. A feed forward neural network is proposed for recognition of digits. 750 digit samples are collected from 15 writers; each writer contributed each of the 10 digits 5 times. Thus a local database containing 750 digit samples is used for training and testing of the proposed OCR system. Preprocessing of handwritten digits is also done before their classification. The combination of proposed feature extraction method along with back-propagation neural network classifier is found to be very effective as it yields excellent recognition accuracy.

1 Introduction

Handwritten digit recognition is a subfield of Optical Character Recognition (OCR) and has always been an active and most fascinating research area in the field of Pattern Recognition and Image Processing for the last few decades. OCR involves the conversion of scanned character/digit images into text that can be edited. In other words, OCR automation is basically the simulation of human reading process and it contributes immensely in numerous applications to improve the man-machine interface.

Off-line handwriting recognition is a subject of much attention due to the presence of many difficulties such as variation in shapes and writing styles of different writers, presence of skew and slant in the handwriting, segmentation of words into characters/digits, different sizes of characters/digits written by different writers etc. Since handwriting depends on the writer and even a single writer can not always write the same character/digit in exactly the same way. As a result, in spite of a dramatic boost in this field of research, off-line handwriting recognition remains an open problem and continues to be an active area for research towards building a recognition system by exploring new techniques and methodologies that would improve recognition performance (accuracy and speed).

An unconstrained handwritten digit recognition system can be divided into several stages such as preprocessing, feature extraction, classification and validation. During the preprocessing stage, various steps are carried out to shape the input digit image into a form suitable for feature extraction [1].

Feature extraction aims at reducing the dimension of input data while extracting relevant information. The feature extraction methods [2] that are widely used are: Template Matching, Zoning, Contour Features, Gabor Features, Gradient Features, Spline Curve, Tangent Features, Box Approach, Fourier Descriptors, Graph Description, Structural Features, Directional Features, Projection Histograms and Randon Transform. Two or more such type of features can be merged as well as selected subset of a particular feature set can be applied to a classifier[3].

In the case of off-line recognition system, an Artificial Neural Network became very popular in the 80s and emerged as fast and most reliable classifier tool resulting in excellent recognition accuracy. In this paper, projection profile features (vertical, horizontal, left diagonal and right diagonal) from the digit image are extracted after preprocessing every digit image and are used to train a feed forward back propagation neural network selected for performing recognition task. Simulation studies are examined extensively and the proposed handwritten digit recognition system is found to deliver excellent recognition performance.

The outline of the paper is as follows. Section 2 briefs the work already done in this field so far. In Section 3, various steps in the proposed recognition system are discussed. Section 4 devotes to sample preparation and preprocessing techniques. Section 5 exposes the feature extraction technique adopted in this work. Section 6 describes the recognition using feed forward back propagation neural network classifier. Discussion of results is presented in section 7 and finally, the paper is concluded in section 8.

2 Brief Review

In the literature, many research articles are available in which researchers have contributed towards the advancements of the automation of handwriting recognition process.

G. S. Lehal and Nivedita Bhatt [4] have proposed a system to recognize both Devnagari and English handwritten numerals. They used a set of global and local features, which were derived from the left and right projection profiles of the numeral image. They tested their system on both Devnagari and English numerals independently and found that recognition rate for Devnagari numerals were better. For Devnagari numerals they found recognition rate of 89% and confusion rate of 4.5%. For English set they found recognition rate of 78.45 and confusion rate of 18%. Rajashekaradhy [5] presented zone centroid and image centroid based distance metric feature extraction system for Indian script numeral recognition. The numeral centroid is computed and the numeral image is divided into n equal zones. Average distance from centroid to the each pixel in the zone is computed. U Bhattacharya and B. B. chaudhuri [6] have promised majority voting scheme for multi-resolution recognition of hand printed numerals. They used the features based of the wavelet transforms at different resolution levels and multilayer perceptron for classification purpose. They achieved 97.165 recognition rates. Faisal Farooq et. al. [7] defined the role of pre-processing in the handwritten recognition. They say that to improve the readability and the automatic recognition of handwritten document images, preprocessing steps are imperative.

3 Proposed Recognition System

Various steps involved in our handwritten digit recognition system are illustrated in Fig.1.

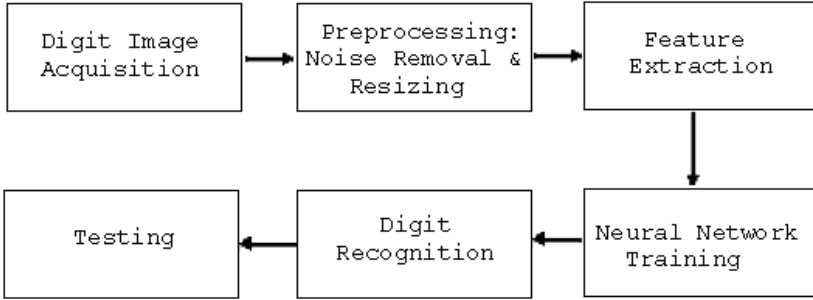


Fig. 1. Schematic diagram of the proposed recognition system

4 Image Acquisition and Sample Preparation

In image acquisition, the digit images are acquired through a scanner or a digital camera. The input digit images are saved in JPEG or BMP formats for further processing. A dataset of English handwritten digits 0-9 is collected from 15 different persons. Each writer contributed 5 samples of each digit. Some of these samples are written on white paper and others on a colored or a noisy background. Some digit samples from our collected dataset are shown in Table 1.

Table 1. Samples contributed by a writer

Digit	Digit Samples				
0	0	0	0	0	0
1	1	1	1	1	1
2	2	2	2	2	2
3	3	3	3	3	3
4	4	4	4	4	4
5	5	5	5	5	5
6	6	6	6	6	6
7	7	7	7	7	7
8	8	8	8	8	8
9	9	9	9	9	9

4.1 Preprocessing

While scanning the digit image, quality of the image is degraded due to the noise that is introduced. It is necessary to remove the background noise to improve the quality of the digit image for our recognition system. The contrast adjustment is also necessary to overcome the problem due to the use of pens of different colors and different intensities of the black color.

Further, as the digits are written in different sizes, it is important to put all the handwritten digit images in a uniform size. To make all the digit image samples in the normal form, all the digit images are reconstructed in the size of 16x16 pixels. Fig. 2(b) shows the resulting image after background noise removal and binarization using grey scale intensity threshold and Fig. 2(c) shows the digit image after resizing.

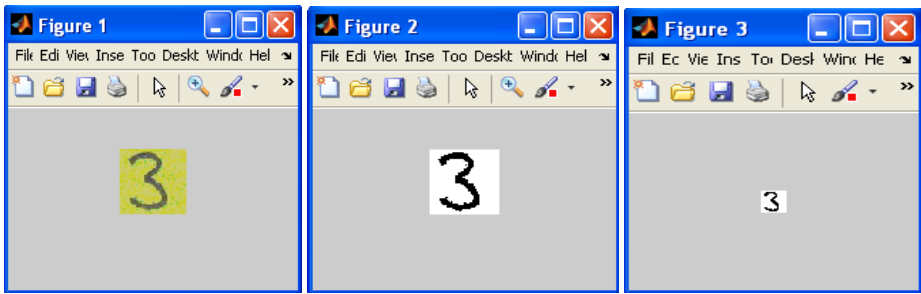


Fig. 2. (a) Original Scanned Image (b) Image after background noise removal and contrast adjustment (c) Resized Image

5 Feature Extraction

Projection profile based feature extraction method delivers excellent results even in the absence of some important preprocessing steps such as smoothing and thinning. In fact, in this type of feature extraction method, it will be disadvantageous to apply the thinning process because there will be a huge loss of important information related to the count and position of black pixels present in the original scanned digit image.

For extracting the features, projection profile of the digit image is computed in vertical, horizontal, left diagonal and right diagonal orientations [8] as indicated in Fig. 3.

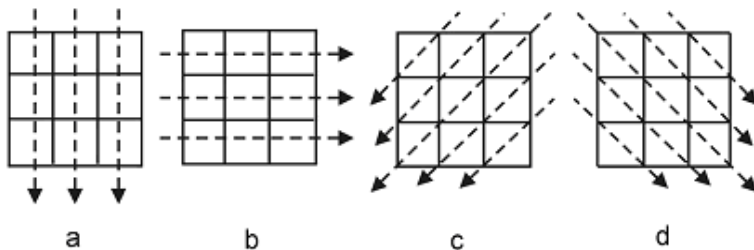


Fig. 3. Pattern Profiles of a 3x3 pattern matrix

Projection profile in vertical direction is computed by scanning the digit image column wise along the y-axis and counting the number of black pixels in each column. As the entire digit images are resized into 16×16 pixels, there are 16 columns and 16 rows. Hence, the vertical projection profile will contain 16 values, each value representing the sum of number of all black pixels present in that particular column.

Similarly, for horizontal projection profile, digit image is traced horizontally along the x-axis. The row wise sum of number of black pixels present in each row will constitute the 16 values of horizontal projection profile. Left diagonal projection profile is computed by traversing the digit image along the left diagonal as shown in Fig. 4(c). The black pixels are counted in left diagonal direction and the sum of the number of black pixels for each left diagonal line of traversing generates 31 values representing the left diagonal projection profile. In the same way, 31 values representing the right diagonal projection profile are obtained by adding the number of black pixels in each right diagonal line of traversing as shown in Fig. 4(d).

Projection profile features in vertical, horizontal, left diagonal and right diagonal orientations to represent a digit ‘3’ are shown in Fig. 4.

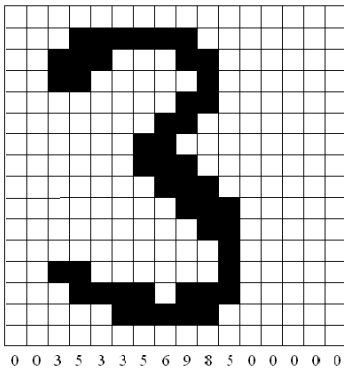


Fig. 4. (a) Vertical profile

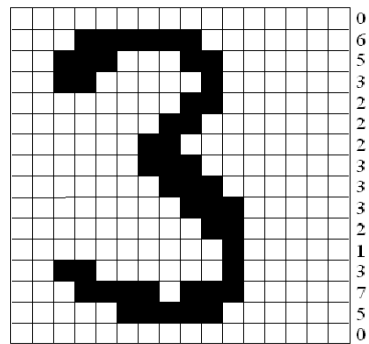


Fig. 4. (b) Horizontal profile

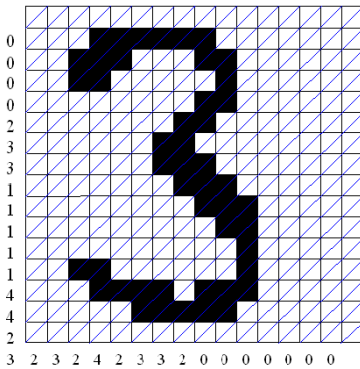


Fig. 4. (c) Left diagonal profile

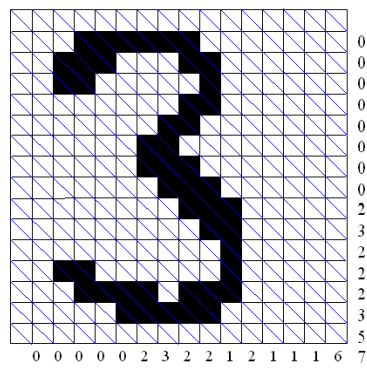


Fig. 4. (d) Right diagonal profile

Projection profile values obtained in all the four orientations (vertical, horizontal, left diagonal and right diagonal) are combined to form a single feature vector. The length of this feature vector is 94 (16+16+31+31=94). A feature vector representing the digit '3' is obtained by concatenation of these four set of features and can be shown as:

```
{0 0 3 5 3 3 5 6 9 8 5 0 0 0 0 0
  0 6 5 3 2 2 2 3 3 3 2 1 3 7 5 0
  0 0 0 0 2 3 3 1 1 1 1 1 4 4 2 3 2 3 2 4 2 3 3 2 0 0 0 0 0 0 0
  0 0 0 0 0 0 0 0 2 3 2 2 2 3 5 7 6 1 1 1 2 1 2 2 3 2 0 0 0 0 0 }
```

6 Classification and Recognition

An extensive review of the literature indicates that as far as the unconstrained handwritten digit recognition is concerned, neural networks as a classifier are chosen to be the best among the others.

6.1 Neural Network Architecture

In the proposed problem of handwritten digit recognition, a feed forward neural network with one hidden layer and employing back propagation algorithm is used for the recognition task. The activation function for neurons of input layer is linear and for hidden and output layer neurons is logsig.

The input layer has 94 neurons as it accepts the feature vector of length 94 representing the input digit image. The output layer has 10 neurons because the network is used to classify 10 digits (0 to 9). The number of neurons in the hidden layer is obtained by trial and error and is kept 50 for optimal results. Performance goal is 0.001 and the maximum allowed epochs count is 100000.

6.2 Neural Network Training

For this experiment, a total of 750 responses are taken into consideration. Features are extracted for all these digit images. The feature vectors of size 94 are applied to the neural network for training. The output is a matrix of size 10x10 because each digit has 10x1 output vector. First column stores the first digit's recognition output; the following column will be for next digit and so on for 10 digits. For each digit, the 10x1 vector will contain value '1' at only one place. For example digit '0' if correctly recognized, will result in [1, 0, 0, 0 ...all ...0] and digit '1' will result in [0, 1, 0, 0 ... all ...0]. The recognition results obtained for various digits are displayed in the form of confusion matrix in Table 2. This confusion matrix shows the confusion among the recognized digits while testing the neural network for the training sample sets.

Table 2. Confusion matrix representing the performance of the classifier

Digit	0	1	2	3	4	5	6	7	8	9	Success (%)
0	73	0	0	0	0	0	1	0	0	1	97.33
1	0	74	0	0	0	0	0	1	0	0	98.66
2	0	0	74	1	0	0	0	0	0	0	98.66
3	0	0	0	72	0	1	0	0	2	0	96.00
4	0	1	0	0	70	0	0	3	0	1	93.33
5	0	0	0	2	0	66	0	0	7	0	88.00
6	1	1	0	0	0	0	69	0	4	0	92.00
7	0	0	0	0	2	0	0	73	0	0	97.33
8	0	0	0	1	0	0	3	0	70	1	93.33
9	0	0	0	0	1	0	0	0	0	74	98.66
Overall Recognition Accuracy											95.33

7 Discussion of Results

The network is trained with 75 sets of each digit i.e 750 digit samples (collected from 15 writers) are there in our training database. The testing of the network is done on the same database used for training. The confusion among the different digits is explained in Table 2. Digit 0 is presented 75 times to the neural network and is classified 73 times correctly. Digit 0 is miss-classified one time as 6 and one time as 9. Digit 1 is misidentified as 7 one time out of seventy five trials. Digit 7 is misclassified as 4 two times. Recognition accuracy for each digit (0-9) as well as overall recognition accuracy is displayed in the last column of Table 2. The average recognition accuracy of around 95% is very good for this handwritten digit recognition work.

8 Conclusions and Future Directions

The use of projection profile features in all the four directions along with back propagation feed forward neural network yielded the excellent recognition accuracy. Although the success rate of 95% is considered excellent but it is not up to the expectations as the testing was done on the training data itself. The performance of a recognition system mainly depends on the quality of samples used for training and the techniques employed to extract the features and the type of classifier. Preprocessing techniques, feature extraction techniques and the methodology used to select the neural network parameters can be improved to get more promising results.

Acknowledgments. The authors acknowledge their sincere thanks to the management and director of Maharaja Surajmal Institute, C-4, Janakpuri, New Delhi, India, for providing infrastructure and financial assistance to carry out this research. The excellent cooperation of the fellow colleagues and library staff is highly appreciated.

References

1. Mantas, J.: An overview of character recognition methodologies. *Pattern Recognition* 19(6), 425–430 (1986)
2. Jain, A.K., Taxt, T.: Feature extraction methods for character recognition-A Survey. *Pattern Recognition* 29(4), 641–662 (1996)
3. Guyon, I., Elisseeff, A.: An Introduction to variable and feature selection. *J. Mach. Learn. Res.* 3, 1157–1182 (2003)
4. Lehal, G.S., Bhatt, N.: A recognition system for Devnagari and English Handwritten Numerals. In: *Proc. ICML*, pp. 442–449. Springer, Heidelberg (2000)
5. Rajashekaradhy, S.V.: Effective zone feature extraction algorithm for handwritten numerals recognition of four south Indian scripts. *Journal of Theoretical and Applied Information Technology* (2008)
6. Bhattacharya, U., Chaudhuri, B.B.: A Majority voting scheme for multiresolution recognition of handprinted numerals. In: *Seventh International Conference on Document Analysis and Recognition (ICDAR)*, vol. 1, p. 16 (2003)
7. Farooq, F., Govindaraju, V., Perrone, M.: Pre-processing Methods for Handwritten Arabic Documents. In: *Proc. of the 8th IEEE International Conference on Document Analysis and Recognition*, vol. 1, pp. 267–271 (2005)
8. Desai, A.A.: Gujarati Handwritten Numeral Optical Character Recognition through Neural Network. *Pattern Recognition* 43(7), 2582–2589 (2010)

Reduct Generation by Formation of Directed Minimal Spanning Tree Using Rough Set Theory

Asit Kumar Das¹, Shampa Sengupta², and Saikat Chakrabarty³

^{1,3} Dept. of Computer Sc. & Tech., Bengal Engineering & Science University, Shibpur,
Howrah – 711 103, West Bengal, India

² Dept. of Information Technology, MCKV Institute of Engineering, Liluah,
Howrah – 711 204, West Bengal, India

akdas@cs.becs.ac.in, shampa2512@yahoo.co.in,
saikatchakrabarty187@gmail.com

Abstract. In recent years, dimension of datasets has increased rapidly in many applications which bring great difficulty to data mining and pattern recognition. Also, all the measured variables of these high-dimensional datasets are not relevant for understanding the underlying phenomena of interest. In this paper, firstly, similarities among the attributes are measured by computing similarity factors based on relative indiscernibility relation, a concept of rough set theory.

Based on the similarity factors, attribute similarity set $AS = \{(A \xrightarrow{k} B) / A, B \text{ are attributes and } B \text{ similar to } A \text{ with similarity factor } k\}$ is formed which helps to construct a directed weighted graph with weights as the inverse of similarity factor k . Then a minimal spanning tree of the graph is generated, from which iteratively most important vertex is selected in reduct set. The iteration completes when the edge set is empty. Thus the selected attributes, from which edges emanate, are the most relevant attributes and are known as reduct. The proposed method has been applied on some benchmark datasets and the classification accuracy is calculated by various classifiers to demonstrate the effectiveness of the method.

1 Introduction

Feature selection and reduct generation are frequently used as a pre-processing step to data mining and knowledge discovery. It selects an optimal subset of features from the feature space according to a certain evaluation criterion. It has been a fertile field of research and shown very effective in removing irrelevant and redundant features, increasing efficiency in data analysis like clustering and classification techniques. In recent years, dimension of datasets has increased rapidly in many applications which bring great difficulty to data mining and pattern recognition. Also, all the measured variables of these high-dimensional datasets are not relevant for understanding the underlying phenomena of interest. This enormity may cause serious problems to many machine learning algorithms with respect to scalability and learning performance. Therefore, feature selection and reduct generation become very necessary for data analysis when facing high dimensional data nowadays. However,

this trend of enormity on both size and dimensionality also poses severe challenges to reduct generation algorithms. Rough Set Theory (RST) [1, 2] is popularly employed to evaluate significance of attributes and helps to find the reduct. The main advantage of rough set theory in data analysis is that it does not need any preliminary or additional information about data like probability in statistics [7], basic probability assignment in Dempster-Shafer theory [8], grade of membership or the value of possibility in fuzzy set theory [9] and so on. But finding reduct by exhaustive search of all possible combinations of attributes is an NP-Complete problem and so many researchers [3-6] applied some heuristic approach for discretization and attribute reduction of real-valued attributes in feature selection.

In the paper, a novel reduct generation method is proposed combining the concept of relative indiscernibility relation [1] of RST and Minimal Spanning Tree (MST) [10]. Relative indiscernibility relation induces partitions of objects from which degree of similarity or similarity factor between two attributes is measured and an attribute similarity (AS) set is obtained. Now, the attribute similarities of AS with similarity factor less than average similarity value are removed and a directed weighted graph is constructed based on the reduced AS set, where weight of an edge is the inverse of the corresponding similarity factor. A minimal spanning tree is obtained from the directed graph using [11]. The tree represents all important similarities of attributes by its edges which help to find out all the information-rich attributes (i.e., vertices) that form the reduct of the data set. To generate reduct, a root (which has no incoming edge) of the spanning tree is selected first and all its outgoing edges are removed. Then another vertex of the maximum out-degree is selected and associated outgoing edges are removed. This process continues until the edge set of the tree becomes empty and all the selected vertices form a reduct.

The rest of the paper is organized as follows: Similarity measurements of attributes and subsequently reduct generation are demonstrated in section 2. Section 3 shows the experimental result of the proposed method and finally conclusion of the paper and the areas for further research are stated in section 4.

2 Proposed Work

The proposed method computes relative indiscernibility of the conditional attributes relative to the decision attribute which helps to measure the degree of similarity among the condition attributes. Based on the similarity of attributes a weighted directed graph is formed and a minimal spanning tree of the graph is obtained which finally generates the reduct.

2.1 Relative Indiscernibility and Dependency of Attributes

Let $DS = (U, A, C, D)$ be a decision system where U is the finite, non-empty set of objects and $A=C \cup D$ such that C and D are set of condition and decision attributes respectively. Each attribute $a \in A$ can be defined as a function, described in (1).

$$f_a: U \rightarrow V_a, \forall a \in A \tag{1}$$

Where, V_a , the set of values of attribute a , is called the *domain* of a .

For any $P \subseteq A$, there exists a binary relation $IND(P)$, called *indiscernibility relation* and is defined in (2).

$$IND(p) = \{(x, y) \in U \times U | \forall a \in p, f_a(x) = f_a(y)\} \tag{2}$$

Where, $f_a(x)$ denotes the value of attribute a for object x in U . Obviously $IND(P)$ is an equivalence relation which induces equivalence classes. The family of all equivalence classes of $IND(P)$, i.e., partition determined by P , is denoted by $U/IND(B)$ or simply U/P and an equivalence class of U/P , i.e., block of the partition U/P , containing x is denoted by $P(x)$.

In the paper, relative indiscernibility relation is introduced based on the concept of conventional indiscernibility relation. It gives indiscernibility of objects for an attribute, relative to another attribute (decision attribute in this case). Every conditional attribute A_i of C determines a relative (to decision attribute) indiscernibility relation (RIR) over U and is denoted as $RIR_D(A_i)$, which can be defined by equation (3).

$$RIR_D(A_i) = \{(x, y) \in \Pi_{A_i}[x]_D \times \Pi_{A_i}[y]_D | f_{A_i}(x) = f_{A_i}(y) \forall [x]_D \in U/D\} \tag{3}$$

where, $\Pi_{A_i}[x]_D$ is the projection operation that selects only the conditional attribute A_i for the objects $[x]_D$, $f_{A_i}(x)$ and $f_{A_i}(y)$ are computed using (1). For each conditional attribute A_i , a relative indiscernibility relation $RIR_D(A_i)$ partitions the set of objects into n -number of equivalence classes, defined as partition $U/RIR_D(A_i)$ or U_D/A_i , equal to $\{\{x\}_{A_i/D}\}$, where $|U_D/A_i| = n$. Obviously, each equivalence class $\{\{x\}_{A_i/D}\}$ contains objects with same decision value which are indiscernible by attribute A_i .

To illustrate the method, a sample dataset represented by Table 1 is considered with eight objects, four conditional and one decision attribute.

Table 1. Sample dataset

Object	Diploma(i)	Experience(e)	French(f)	Reference(r)	Decision
X_1	MBA	Medium	Yes	Excellent	Accept
X_2	MBA	Low	Yes	Neutral	Reject
X_3	MCE	Low	Yes	Good	Reject
X_4	MSc	High	Yes	Neutral	Accept
X_5	MSc	Medium	Yes	Neutral	Reject
X_6	MSc	High	Yes	Excellent	Reject
X_7	MBA	High	No	Good	Accept
X_8	MCE	Low	No	Excellent	Reject

Here, equivalence classes by $IND(P)$ and $RIR_D(A_i)$ are formed using (2) and (3) respectively and listed in Table 2.

Table 2. Equivalence classes by two different relations

Equivalence classes by $IND(P)$	Equivalence classes by $RIR_D(A_i)$
$U/D = (\{x_1, x_4, x_7\}, \{x_2, x_3, x_5, x_6, x_8\})$	$U_D/i = (\{x_1, x_7\}, \{x_2\}, \{x_3, x_8\}, \{x_4\}, \{x_5, x_6\})$
$U/i = (\{x_1, x_2, x_7\}, \{x_3, x_8\}, \{x_4, x_5, x_6\})$	$U_D/e = (\{x_1\}, \{x_5\}, \{x_2, x_3, x_8\}, \{x_4, x_7\}, \{x_6\})$
$U/e = (\{x_1, x_5\}, \{x_2, x_3, x_8\}, \{x_4, x_6, x_7\})$	$U_D/f = (\{x_1, x_4\}, \{x_2, x_3, x_5, x_6\}, \{x_7\}, \{x_8\})$
$U/f = (\{x_1, x_2, x_3, x_4, x_5, x_6\}, \{x_7, x_8\})$	$U_D/r = (\{x_1\}, \{x_6, x_8\}, \{x_2, x_5\}, \{x_4\}, \{x_3, x_7\})$
$U/r = (\{x_1, x_6, x_8\}, \{x_2, x_4, x_5\}, \{x_3, x_7\})$	

2.2 Formation of Attribute Similarity Set Using Similarity Measurement

An attribute A_i is similar to another attribute A_j in context of classification power if they induce the same equivalence classes of objects under their respective indiscernibility relations. But in real situation, it rarely occurs and so similarity of attributes is measured by introducing the similarity measurement factor which indicates the degree of similarity of one attribute to another attribute. Here, an attribute A_i is said to be similar to an attribute A_j with degree of similarity (or similarity factor) $\delta_f^{i,j}$ and is denoted by $A_i \xrightarrow{\delta_f^{i,j}} A_j$ if the probability of inducing the same equivalence classes of objects under their respective relative indiscernible relations is $(\delta_f^{i,j} \times 100)\%$, where $\delta_f^{i,j}$ is computed by equation (4).

$$\delta_f^{i,j} = \frac{1}{|U_D/A_i|} \sum_{[x]_{A_i/D} \in U_D/A_i} \frac{1}{|[x]_{A_i/D}|} \max_{[x]_{A_j/D} \in U_D/A_j} \left(|[x]_{A_i/D} \cap [x]_{A_j/D} \right) \quad (4)$$

It is quite obvious that $\delta_f^{i,j}$ would have value 1 if A_i and A_j have exactly similar classification pattern. For each pair of conditional attributes (A_i, A_j) , similarity factor is computed by (4). High value of similarity factor of $A_i \rightarrow A_j$ means that the relative indiscernibility relations $RIR_D(A_i)$ and $RIR_D(A_j)$ produce highly similar equivalence classes. This implies that both the attributes A_i and A_j have almost similar classification power and so $A_i \rightarrow A_j$ is considered as strong similarity of A_i to A_j . Since, for any two attributes A_i and A_j , two similarities $A_i \rightarrow A_j$ and $A_j \rightarrow A_i$ are obtained, only one with higher similarity factor is selected in the list of attribute similarity set AS. Thus, for n attributes, $\frac{n(n-1)}{2}$ similarities are selected, out of which some are strong and some are not. Out of these, the similarities with $\delta_f^{i,j}$ value less than the average δ_f value are discarded and rest is considered as the set of attribute similarity AS. So, each element x in AS is of the form $x: A_i \rightarrow A_j$ such that $\text{Left}(x)=A_i$ and $\text{Right}(x)=A_j$. The algorithm ‘‘ASS_GEN’’ described below, computes the attribute similarity set AS.

```

Algorithm: ASS_GEN(C,  $\delta_f$ )
/* Computes attribute similarity set  $\{A_i \rightarrow A_j\}$  */
Input: C = set of conditional attributes and  $\delta_f = 2$ -D matrix
containing similarity factors between each pair of conditional
attributes, obtained using (4).
Output: Attribute Similarity Set AS
Begin

AS = {}, sum_ $\delta_f = 0$ ;
/* compute only n(n - 1)/2 elements in AS */
for i = 1 to |C| - 1 {
  for j = i+1 to |C| {
    if ( $\delta_f^{i,j} > \delta_f^{j,i}$ ) {sum_ $\delta_f = \text{sum\_}\delta_f + \delta_f^{i,j}$ ;
      AS = AS  $\cup \{A_i \rightarrow A_j\}$ }
    else {sum_ $\delta_f = \text{sum\_}\delta_f + \delta_f^{j,i}$ ; AS = AS  $\cup \{A_j \rightarrow A_i\}$ }
  } /* end of i and j loops */
}
/* modify AS to store only  $\{A_i \rightarrow A_j\}$  for which  $\delta_f^{i,j} > \text{avg\_}\delta_f$  */
ASmod = {}; avg_ $\delta_f = \frac{2 \times \text{sum\_}\delta_f}{|C|(|C|-1)}$ ;
for each  $\{A_i \rightarrow A_j\} \in \text{AS}$  { if ( $\delta_f^{i,j} > \text{avg\_}\delta_f$ ) {
  ASmod = ASmod  $\cup \{A_i \rightarrow A_j\}$ ; AS = AS -  $\{A_i \rightarrow A_j\}$ 
}
}
AS = ASmod
End.
    
```

Initially, algorithm “AS_GEN” selects $\text{AS} = \{i \rightarrow f, i \rightarrow r, e \rightarrow i, e \rightarrow f, e \rightarrow r, r \rightarrow f\}$ and constructs Table 3. As the average similarity factor $\text{avg_}\delta_f = 0.786$ which is less than the similarity factors for attribute similarities $i \rightarrow f, e \rightarrow i, e \rightarrow f$ and $r \rightarrow f$, the modified attribute similarity set $\text{AS} = \{i \rightarrow f, e \rightarrow i, e \rightarrow f, r \rightarrow f\}$.

Table 3. Selection of attribute similarities in AS

Attribute Similarity ($A_i \rightarrow A_j; i \neq j$ and $\delta_f^{i,j} > \delta_f^{j,i}$)	Similarity Factor of $A_i \rightarrow A_j$ ($\delta_f^{i,j}$)	$\delta_f^{i,j} > \delta_f$
$i \rightarrow f$	$\delta_f^{i,j} = 0.8$	Yes
$i \rightarrow r$	$\delta_f^{i,r} = 0.7$	
$e \rightarrow i$	$\delta_f^{\theta,i} = 0.83$	Yes
$e \rightarrow f$	$\delta_f^{\theta,f} = 0.83$	Yes
$e \rightarrow r$	$\delta_f^{\theta,r} = 0.76$	
$r \rightarrow f$	$\delta_f^{r,f} = 0.8$	Yes
Average δ_f	0.786	

2.3 Minimal Spanning Tree Generation of Attribute Similarity Graph

The minimized attribute similarity set $AS = \left\{ A_i \xrightarrow{\delta_f^{i,j}} A_j (A_i \neq A_j) \right\}$ contains the set of pairs of attributes that are most strongly related to each other. To generate a reduct, firstly this set is represented by a directed graph, called *attribute similarity graph* (ASG). The vertices of ASG are the conditional attributes present in the set AS and weighted edge exists from attribute A_i to attribute A_j with weight $\delta_f^{i,j}$ if $A_i \xrightarrow{\delta_f^{i,j}} A_j \in AS$. Thus, attribute similarity $A_i \rightarrow A_j$ with $\delta_f^{i,j} = w$, present in set AS is represented by a directed edge from vertex A_i to vertex A_j with weight w . The ASG, therefore, represents the total similarity structure of the similarity set AS. Some vertices in the ASG may have multiple incoming edges which imply that a particular vertex v is similar to more than one other vertex. Without loss of generality, if one of these vertices to which v is the most similar can be identified, the other edges incident on v may be dropped. To construct the minimal spanning tree, weights associated to each edge of the directed graph ASG are inversed and Chu-Liu/Edmond's Algorithm [11] is applied. In the process, the vertices that have only outgoing edges and no incoming edges are considered as the good candidates for the selection of a root. If more than one such vertex exists, then they are fused to form a single vertex. So, before construction of the minimal spanning tree, ASG is modified to merge all the nodes with in-degree zero to a single node and considered it as the root of the graph.

Algorithm: MST_GEN(AS)

```
/* generates minimal spanning tree of ASG */
Input: AS = modified attribute similarity set obtained from
ASS_GEN algorithm.
Output: Rooted Directed Minimal Spanning Tree M
Begin
```

```
Construct weighted graph ASG = (V, E) from AS, where
V = {Ai | Ai ∈ Left(x) ∪ Right(x), ∀x ∈ AS}
```

$$E = \left\{ (A_i, A_j) \mid A_i \xrightarrow{\delta_f^{i,j}} A_j \in AS \right\}$$

```
/* Merge nodes with in-deg zero to create a new node */
```

```
Root = { }
```

```
for each vertex Ni ∈ V {if(in_deg(Ni) = 0){
```

```
    Root = Root ∪ {Ni}
```

```
    Modify ASG by fusing all vertices in Root}}
```

```
for each edge Ai  $\xrightarrow{\delta_f^{i,j}}$  Aj ∈ E { $\delta_f^{i,j} = (\delta_f^{i,j})^{-1}$ }
```

```
/* Compute MST of ASG using Chu-Liu/Edmond's algorithm */
```

```

for each vertex  $v \in V - \text{Root}$ 
select the entering edge with the smallest cost;
Let  $S =$  selected  $|V - \text{Root}|$  edges;
Repeat {
  If (no cycle)  $\text{MST}(V, S)$  is a minimal Spanning Tree;
  Else {for each cycle formed {
    Merge vertices in cycle to a new vertex ( $k$ );
    Modify the cost of each edge which enters a vertex( $j$ )
in the cycle from some vertex( $i$ ) outside the cycle using
 $c(i, k) = c(i, j) - (c(x(j), j) - \min_{\{j\}}(c(x(j), j)))$ , where  $c(x(j), j)$ 
is the cost of the edge in the cycle which enters  $j$ ;}
  For each new vertex {
    Select the entering edge with smallest modified cost
    Replace the existing edge by the new selected edge}
Until (no cycle formed);
End.

```

The attribute similarity graph (ASG) generated from set AS, modified ASG and corresponding minimal spanning tree (MST) are shown in Fig. 1(a), Fig. 1(b) and Fig. 1(c) respectively.

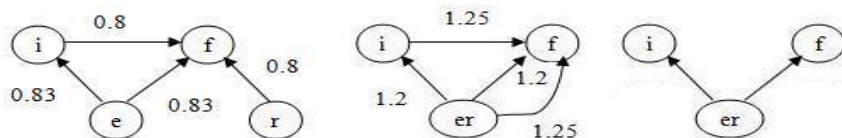


Fig. 1. (a) ASG

(b) Modified ASG

(c) MST of Fig. 1(b)

2.4 Reduct Generation

The above generated rooted directed minimal spanning tree would give the highest similarities between the attributes. In the final stage, the maximal spanning tree is searched to find the vertex with highest out-degree. The vertex with highest out-degree is an attribute to which most number of other attributes is similar. So, this node is added to the initially empty reduct set and its out-going edges are removed from the tree. This process of trimming the edges of the tree and adding the vertex (attribute) to the reduct set continues till the edge set of the tree becomes empty and thus final reduct is obtained.

Algorithm: RED_GEN(MST)

/* generates reduct from rooted directed minimal spanning tree of ASG */

Input: $\text{MST}(V, S) =$ Rooted Directed Minimal Spanning Tree

Output: Reduct

Begin

$R = \{ \}$

order[V]= array of vertices of MST sorted in descending order of their out-degree

```

for i = 1 to |V| {
  Remove outgoing edges from vertex order[i]
  R = R ∪ {order[i]}
  if (S = ∅) return (R)}
End

```

Reduct generated from Fig. 1(c) is {e, r} as shown in Fig. 2.

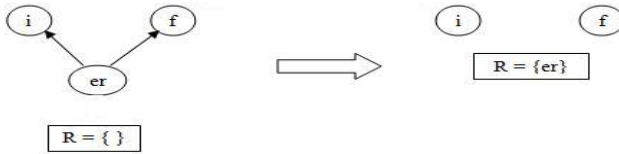


Fig. 2. Reduct Generation from Minimal Spanning Tree

3 Experimental Results

The proposed method computes a single reduct for datasets collected from UCI machine learning repository [12]. To measure the efficiency of the method, k-fold cross-validations, where k ranges from 1 to 10 have been carried out on the dataset and classified using “Weka” tool [13]. The proposed method (PRP) and well known dimensionality reduction methods, such as, *Cfs Subset Evaluation* (CFS) method [14] and *Consistency Subset Evaluator* (CON) method [15] have been applied on the dataset for dimension reduction and the reduced datasets are classified on various classifiers. Original number of attributes, number of attributes after applying various reduction methods and the accuracies (in %) of the datasets are computed and listed in Table 4, which shows the efficiency of the proposed method.

Table 4. Accuracy Comparison of Proposed, CFS and CON methods

Classifier	Wine (13)			Heart (13)			Glass (9)		
	PRP (9)	CFS (8)	CON (8)	PRP (9)	CFS (8)	CON (11)	PRP (8)	CFS (7)	CON (9)
Naïve Bayes	93.70	94.80	95.30	83.27	84.38	85.50	67.28	43.92	47.20
SMO	94.91	94.30	93.74	83.27	84.75	80.38	64.48	57.94	57.48
KSTAR	95.48	92.17	93.17	83.81	82.15	81.78	83.64	79.91	78.50
Bagging	92.09	90.35	90.91	82.52	82.52	83.64	76.63	73.83	71.50
J48	92.65	92.17	92.61	82.89	80.52	81.15	70.09	68.69	64.20
PART	92.09	90.17	91.17	79.43	81.41	78.55	75.23	70.09	68.60
Average	92.64	92.30	92.80	82.53	82.60	81.80	68.50	63.20	62.10

4 Conclusion and Future Enhancements

The paper describes a new method of attribute reduction using minimal spanning tree. It does not use any heuristic algorithm which gives good result only if the heuristic is powerful. The results show that the new method is good enough and often gives better accuracy than the existing ones in most of the cases. Future enhancements to this work may include generation of all possible maximal spanning trees to compute multiple reduct sets and finally select the best one for classification.

References

1. Pawlak, Z.: Rough sets. *International Journal of Information and Computer Sciences* 11, 341–356 (1982)
2. Pawlak, Z.: Rough set theory and its applications to data analysis. *Cybernetics and Systems* 29, 661–688 (1998)
3. Hu, X., Lin, T.Y., Jianchao, J.: A New Rough Sets Model Based on Database Systems. *Fundamental Informaticae*, 1–18 (2004)
4. Jensen, R., Shen, Q.: Fuzzy-Rough Attribute Reduction with Application to Web Categorization. *Fuzzy Sets and Systems* 141(3), 469–485 (2004)
5. Zhong, N., Skowron, A.: A Rough Set-Based Knowledge Discovery Process. *Int. Journal of Applied Mathematics and Computer Science* 11(3), 603–619 (2001); *BIME Journal* 05(1) (2005)
6. Kerber, R.: ChiMerge: Discretization of Numeric Attributes. In: *Proceedings of AAAI-1992, Ninth Int'l Conf. Artificial Intelligence*, pp. 123–128. AAAI Press (1992)
7. Devroye, L., Györfi, L., Lugosi, G.: *A Probabilistic Theory of Pattern Recognition*. Springer, Newyork (1996)
8. Gupta, S.C., Kapoor, V.K.: *Fundamental of Mathematical Statistics*. Sultan Chand & Sons, A.S. Printing Press, India (1994)
9. Pal, S.K., Mitra, S.: *Neuro-Fuzzy pattern Recognition: Methods in Soft Computing*. Willey, New York (1999)
10. Bang-Jensen, J., Gutin, G.: *Digraphs: Theory, Algorithms and Applications*. Springer, Heidelberg, ISBN 1-85233-268-9
11. Chu, Y.J., Liu, T.H.: On the shortest arborescence of a directed graph. *Science Sinica* 14, 1396–1400 (1965)
12. Murphy, P., Aha, W.: UCI repository of machine learning databases (1996), <http://www.ics.uci.edu/mllearn/MLRepository.html>
13. WEKA: Machine Learning Software, <http://www.cs.waikato.ac.nz/~ml/>
14. Hall, M.A.: *Correlation-Based Feature Selection for Machine Learning* PhD thesis, Dept. of Computer Science, Univ. of Waikato, Hamilton, New Zealand (1998)
15. Liu, Setiono, R.: A Probabilistic Approach to Feature Selection: A Filter Solution. In: *Proc. 13th Int'l Conf. Machine*

Misclassification and Cluster Validation Techniques for Feature Selection of Diseased Rice Plant Images

Santanu Phadikar¹, Asit Kumar Das², and Jaya Sil²

¹ Department of Computer Science and Engineering, West Bengal University of Technology, Saltlake, Kolkata -64, India

² Department of Computer Science and Technology, Bengal Engineering and Science University, Shibpur Howrah, India

sphadikar@yahoo.com, akdas72@rediffmail.com

Abstract. Damages of rice crops can be averted by taking corrective measures at an early stage based on the classification of diseases of rice plants. The paper aims at developing an appropriate data mining methodology to extract knowledge about the characteristics of diseases by analyzing the images acquired from the field. Since all features are not equally involved in classifying diseases; selection of optimum features is a challenging task to address the problem. The work is performed in three steps. Firstly, thirty six features of different category are extracted from the diseased plant images using image processing techniques. Then, images with respect to each attribute are clustered by k-means algorithm where k is the number of class labels of the diseases. Cluster validity indices are computed using the proposed and existing techniques. Finally, all the computed indices are fused and based on that score of the attributes is calculated. The attributes having scores greater than average score are considered as optimal feature set. With the reduced feature set, classification accuracy is calculated by applying the proposed method on four hundred fifty infected rice plant images and compared with other classifiers demonstrating effectiveness of the proposed model.

Keywords: Rice Diseases, Image Processing, Feature Selection, Cluster Validations, Misclassification.

1 Introduction

With the advancement of cost effective computer and internet technology, application of technology becomes an integral part in our daily life, including agro business [1], weather forecasting [2], GIS based production and damage measurement [3] and precision agriculture [4-6]. Precision agriculture concentrates on providing the means for observing, assessing and controlling agricultural practices in order to develop an automated system of crop management. It also takes into account the pre- and post-production aspects of agricultural enterprises. The objectives of precision agriculture are profit maximization, agricultural input rationalization and environmental damage reduction, by adjusting the agricultural practices to the site demands. The challenge of the precision approach is to equip the farmer with adequate and affordable

information and control technology. Plant disease is one of the crucial causes that reduces quantity and degrades quality of the agricultural products. Rice is second largest crop produce in the World and the first largest crop in India. Thus automatic diagnosis of rice plant diseases plays a significant role not only from researcher's point of view but also from the economical view. One of the most significant branches in precision agriculture is to automatic diagnoses of the field problem at an early stage to minimize the use of pesticides and that maximizes the productivity.

The automatic disease diagnosis system found in the literature [5-6] uses several features like textures [7-8], color, shape and several other features of the images to detect the diseases. However, use of too many features for disease detection makes the system not only complex, but it also compromise with the accuracy of classification due to redundancy and noise in the extracted features. Therefore, selection of optimized feature set by image analysis is an important activity for obtaining better prediction accuracy in diagnosis of diseases classification.

Data mining and soft computing techniques [9-11] are applied to discover knowledge of plant diseases by deriving classification rules comprise of characteristics of diseases and their class labels, reported in [6,10]. Analysis of diseased images reveal the fact that accurate diagnosis depends on the visual properties of the plants such as change of color, shape, orientation (textures) of the infected portion of the images. One of the most important problems of the automatic diagnosis process is to identify the significant information from large volume of data using appropriate data mining techniques. Therefore, feature selection [12-13] has become an important pre-processing step to reduce complexity in building an efficient classifier [14] for diagnosing the diseases.

In the proposed method, different diseased rice plant images acquired from the rice field are used as training dataset to build the classifier. Various types of image features are extracted using image processing techniques and categorized based on colors, shapes and texture. Change of color, deviation from the actual shape and non-uniformity of the infected leaf provide important information to diagnose the diseases. Following steps are performed to select only the important features from the datasets.

(i) Images are clustered with respect to each attribute by k-means algorithm where k is the number of class labels of the diseases. Considering actual class labels of the objects given in the dataset, minimum number of misclassified objects for each cluster is calculated and finally misclassification indices (MI) of the attributes are obtained. Lower the index value of an attribute implies more important the attribute is.

(ii) Corresponding to each attribute, the objects are partitioned into different classes based on their actual class labels. Then, the validity indices such as Dunn's Index (DI) [15], DB index (DB) [16] and CS Index (CS) [17] of clusters (here, classes) are computed by different cluster validation techniques [18]. Lower the DB and CS values and higher the DI value of a cluster imply more important the corresponding attribute is.

(iii) Since, higher the DI value implies more important the corresponding attribute, so lower the inverse of DI value implies more important the attribute is. Now, fusing the inverse of DI value and values of other indices computed indices, score of the attribute is generated and the attributes with less than average score are considered as optimal feature set.

The rest of the paper is organized as follow: Section 2 describes the image acquisition and feature extraction methods. Feature selection process has been described in section 3. Section 4 describes the results of the proposed method and compares it with different feature selection methods to demonstrate the efficiency of the method. Finally, conclusion has been summarized in section 5.

2 Feature Extraction

Plant disease is one of the crucial causes that reduces quantity and degrades quality of the agricultural products. Rice is second largest crop produce in the World and the first largest crop in India. Thus automatic diagnosis of rice plant diseases plays a significant role from environmental and economical point of view. Most common rice plant diseases [19-21] such as *Brown spot*, *Leaf blast*, and *Sheath rot* are considered in the paper to design a classifier with three different class labels. Among these diseases, in open eye classification of *Leaf blast* and *Brown spot* becomes very difficult [20]. Therefore, a computational approach has been proposed to develop an automated system for diagnosing the diseases.

Data for the proposed classification method has been extracted using the following two steps.

1. Image acquisition
2. Segmentation

2.1 Image Acquisition

For the proposed work sample images have been acquired from the rice field of various rice growing districts in West Bengal, India. Acquired images are then verified and vetted by agro scientists for obtaining the disease class labels, considered as actual class label. After acquisition, select only the infected portion surrounded by some normal portion is cropped. Then noises due to dust, spores and water in the images are filtered by applying median filter of size 5×5 to obtain resultant images, shown in Fig. 1.

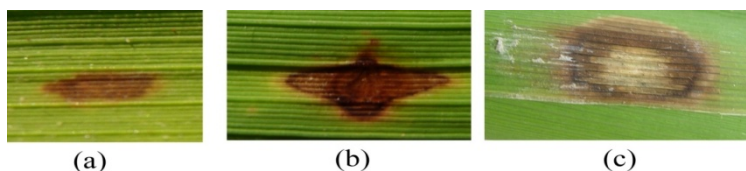


Fig. 1. The preprocessed images (a) Leaf infected by Brown Spot (b) Leaf with Blast diseases (c) Stem with Sheath rot

2.2 Segmentation

Literatures in the field of agriculture [19-21] suggest that the shape of the spots, change of color and the orientation of colors and shapes in the infected portion

(texture) provide important information for classification of diseases. So features are grouped based on, shape, color and texture features to classify the diseases.

In order to extract all these features, infected portion of the spot has been separated from the normal portion. First the color images of RGB (Red Green and Blue) plane is converted into HSV (Hue, Saturation and Value) plane and by applying Otsu's threshold based segmentation [22] in the Hue plane, segmented images are obtained shown in Fig. 2.

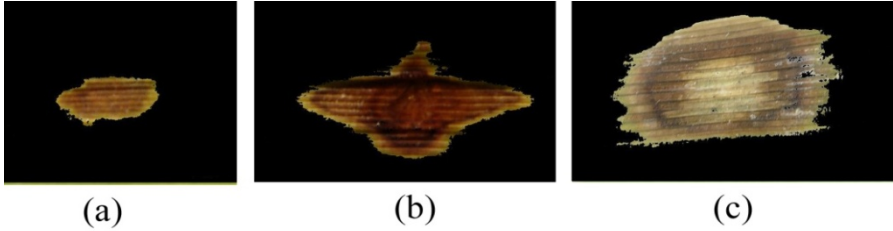


Fig. 2. (a) Segmented brown Spot Image (b) Blast image (c) Sheath rot image

2.2.1 Shape Based Features

It has been observed that the spot created by *Brown spot* disease are oval shaped, elliptical shape is due to *Leaf blast* disease with pointed end and *Sheath rot* disease damages the leaf with irregular spots. Different features which can predict the shape of the spot formed due to different diseases are extracted by applying geometrical computation methods. Area (AR) of the infected spots, Sharpness (SH), Area-discrepancy (AD) deviation of area from the boundary ellipse, Perimeter (PR), Eigen values (EV), Aspect-ratio (ASR) and seven central moments [23] of degree three, invariant to scaling, translation and rotation are computed as shape based features to detect the diseases as listed in Table 1.

2.2.2 Color Based Features

Color of the Center and boundary of the spots are distinct for different diseases. Color features are obtained by calculating mean (M) and standard deviation (SD) of the pixels within the spots, considering three classical planes; red (R), green (G) and blue (B). Color features are extracted from the center of the spots (CR), Boundary of the spots (BR) and back ground (BC) of the segmented images.

All 36 extracted features are listed in Table 1 using their abbreviated name. For example, BC_M_R, and BC_SD_R represent mean (M) and standard deviation (SD) of the spot in the background region (BC) by considering red (R) color plane.

2.2.3 Texture Based Features

Orientation of colors and shapes is represented by the image texture. Different texture features extracted from the diseased images are energy (EG), entropy (ET), contrast (CT), homogeneity (HG) and co-relation (CR), as listed in Table 1.

Table 1. Describe various extracted features of Rice Plant Images

Feature Category	Feature Names
Colour Features	BC_M_R, BC_SD_R, BC_M_G, BC_SD_G, BC_M_B, BC_SD_B, BR_M_R, BR_SD_R, BR_M_G, BR_SD_G, BR_M_B, BR_SD_B, CR_M_R, CR_SD_R, CR_M_G, CR_SD_G, CR_M_B, CR_SD_B
Shape Features	AR, SH, AD, PR, EV, ASR, $\phi_1, \phi_2, \phi_3, \phi_4,$ ϕ_5, ϕ_6, ϕ_7
Texture Features	EG, ET, CT, HG, CR

3 Optimum Feature Selection

Once the features are extracted, the decision table is constructed with 36 features of 450 infected rice plant images and three disease class labels as describe in section 2. To reduce the complexity of classification, proposed method redefines the decision table by selecting only the relevant features without compromising its accuracy. The set of relevant features, called reduct [12-13] determines the optimal set of features for diagnosing the diseases.

First of all, the objects (images) are classified into k number of classes based on their actual class labels in the decision table where $CLASS = \{CLASS_1, CLASS_2, \dots, CLASS_k\}$ and k is the number of class labels. Now for each feature or attribute $A_t \in \{A_1, A_2, \dots, A_n\}$, the images are partitioned into k clusters $CLUS = \{CLUS_1, CLUS_2, \dots, CLUS_k\}$, using k -means clustering algorithm. Hence a one to one correspondence is required between actual classes of the objects and their presence in the clusters so that misclassification of the objects becomes the minimum. So, for each permutation $\langle CLUS_{i_1}, CLUS_{i_2}, \dots, CLUS_{i_k} \rangle$ of clusters in $CLUS$, a one to one correspondence is considered based on their positions and misclassification index (MI) of the text images with respect to attribute A_t is computed using (1). The minimum value of MI is the final misclassification index value of A_t . Lower the index value of an attribute implies more important the attribute is.

$$MI(A_t) = \sum_{j=1}^k \frac{1}{|CLUS_{i_j}|} [CLUS_{i_j} - \{CLUS_{i_j} \cap CLASS_j\}] \quad (1)$$

Corresponding to each attribute, the objects are partitioned into different classes based on their class values. Then, the validity indices such as Dunn's Index (DI), DB index (DB) and CS Index (CS) of clusters are computed by different cluster validation techniques [15-18]. Lower the DB and CS values and higher the DI value of a cluster imply more important the corresponding attribute. Since, all computed indices except DI index are lower for better cluster, so score of the attribute is obtained by averaging MI, DB, CS and inverse of DI index where, lowest score gives most important attribute. Same process is followed to compute the score of all attributes and the attributes with final score greater than the average final score are selected as optimal feature set or reduct.

Detailed algorithm for optimal reduct generation has been given below:

```

Algorithm:: Reduct_Generation( DATA, REDUCT)
/*DATA is a two dimensional matrix of size M ×(N+1)
where M is the Number of images and N is the number of
features and last column contains the decision value.
REDUCT is the selected features */
Begin
  1. Initialize REDUCT = ∅ and K = number of distinct
     values in decision attribute.
  2. Let CLASS = {CLASS1, CLASS2, ..., CLASSk} is the
     classes of objects based on decision values.
  3. Let A = {A1, A2, ..., An} be the feature set.
  4. For I =1 to n
     a. Apply K-mean Clustering on attribute AI and
        obtain clusters CLUS = { CLUS1, CLUS2, ..., CLUSk }
     b. MI(AI) = ∞
     c. For each permutation CL={CLUSi1, CLUSi2, ..., CLUSik} of
        CLUS
        /*Compute misclassification index*/
        (i)  $\delta = \sum_{j=1}^k \frac{1}{|CLUS_{i_j}|} [CLUS_{i_j} - \{CLUS_{i_j} \cap CLASS_j\}]$ 
        (ii) IF ( $\delta < MI(A_I)$ ) THEN MI(AI) =  $\delta$ 
     d. Determine Dunn's Index (DI), DB index (DB)
        and CS Index (CS) of clusters obtained for
        attribute AI.
     e. SCORE(AI)=(MI + CS + DB + (DI)-1)/4
  5. Calculate average score AVG_SCR from SCORE.
  6. For I=1 to n
     IF (SCORE(AI) < AVG_SCR)
       REDUCT = REDUCT ∪ {AI}
  7. Return (REDUCT)
End

```

4 Result and Discussion

The proposed method is applied on a dataset generated from 450 infected rice plant images of three diseased classes (brown spot, blast and sheath rot), with 36 features. Sample datasets for different kinds of features, calculated using the methodologies is described in section 2. For each of 36 features MI, DI, DB and CS indices are calculated and finally, based on their scores 18 features have been selected, which includes {BR_M_G, BR_SD_R, BR_SD_G, BR_SD_B, BC_M_G, BC_DS_R, BC_DS_G, CR_M_R, CR_M_G, CR_M_B, CR_SD_G, CR_SD_R, CR_SD_B, ET, AR, EG, HG, PR}. Thus, the proposed method selects 18 features, whereas, "Cfs Subset Eval" (CFS) method selects 19 features and "Consistency Subset Eval" (CON)

with Rank Search method finds 20 features out of thirty-six extracted features of the disease images. So the rate of dimensionality reduction is higher for the proposed method compare to the existing methods. The method does not reduce dimension of data by losing its decision making capability, rather it provides compatible classification accuracy obtained by various classifiers when run using “weka” tool [24] where 10-fold cross-validations are carried out, as listed in Table 2. In Table 2, other dimension reduction methods like “Principal Component Analysis” (PCA), “Chi-Squared Attribute Eval” (CHI), “Classifier Subset Eval” (CLS) and “Support Vector Machine Attribute Eval” (SVM) are used where first eighteen ranked attributes are considered for classification, as the proposed method selects only eighteen attributes. The accuracy of classifiers shows that the proposed method is at least comparable with other dimensionality reduction methods.

Table 2. Accuracy of Different Classifier for Reduced Dataset

Classifier	Proposed Method	PCA	CHI	CLS	SVM	CFS	CON
C4.5	86.21	83.79	84.6	84.85	84.34	84.85	84.85
PART	87.59	87.77	84.85	84.45	84.34	86.36	87.12
K-STAR	89.63	88.05	91.16	89.65	89.9	86.87	88.89
NaïveBaye’s	85.8	83.74	79.55	80.3	85.61	84.09	80.05
SMO	89.88	88.33	88.89	88.13	90.44	89.39	89.14
Boosting	78.1	74.44	75.25	75.25	75.76	75.5	75.25
Bagging	87.86	86.56	86.11	85.61	85.61	86.62	86.11
MCS	91.65	88.63	90.4	89.5	92.42	90.4	90.66
Average	87.09	85.16	85.10	84.72	86.05	85.51	85.26

5 Conclusion

The proposed method encompasses a novel strategy in dimensionality reduction by attribute clustering based on the classification ability of the individual attribute in the system. To improve the performance of the system well known cluster validation indices such as Dunn’s Index (DI), DB index (DB) and CS Index (CS) are clubbed with the proposed attribute ranked method based on misclassification of images. The rate of dimension reduction of the rice plant image dataset is measured and compared with existing methods as well as the classification accuracy with reduced dataset is calculated by various classifiers to measure the effectiveness of the method.

References

1. Arumapperuma, S.: Role of Information Technology in the Diffusion of Innovations to the Farm Level within the Australian Agribusiness System. In: European Federation of IT in Agriculture and the World Congress on Computers in Agriculture (EFITA/WCCA). Glasgow Caledonian University (2007)

2. Cantelaube, P., Terres, J.-M.: Seasonal weather forecasts for crop yield modelling in Europe. *Tellus A* 57, 476–487 (2005)
3. Qin, Z., Zhang, M., Christensen, T., Li, W., Tang: Remote Sensing Analysis of Rice Disease Stress for Farm Pest Management Using Wide-band Airborne Data. In: *IEEE Geoscience & Remote Sensing Symposium*, pp. 2215–2217 (2003)
4. Wachs, J.P., Stem, H.I., Burks, T., Alchanatis, V.: Low and high-level visual feature-based apple detection from multi-modal images. *Journal of Precision Agriculture* 11(6), 717–735, <http://www.springerlink.com>, doi:10.1007/s11119-010-9198-x
5. Phadikar, S., Sil, J.: *IEEE International Conference on Information Technology Rice Disease Identification using Pattern Recognition Techniques*, pp. 420–423 (2008)
6. Phadikar, S., Sil, J., Das, A.K.: Feature Selection By Attribute Clustering Of Infected Rice Plant Images. *International Journal of Machine Intelligence* 3(2), 74–88 (2011)
7. Bharti, M.H., Liu, J.J., Macgregor, J.F.: Image texture analysis: methods and comparisons. *Chemo Metrics & Intelligence Laboratory Systems* 72, 57–71 (2004)
8. Haralick, R.M.: Statistical and Structural Approaches to Texture. *Proceedings of the IEEE* 67, 786–804 (1979)
9. El Mohammed, T., Mahmoud, W., El Mahmoud, B.: Rice Genome Sequencing and Data Mining Resources. *The International Arab Journal of Information Technology* 3(4), 303–307 (2006)
10. Lu, W., Han, J., Ooi, B.C.: Knowledge Discovery in Large Spatial Databases. In: *Proc. Far East Workshop Geographic Information Systems*, pp. 275–289 (1993)
11. Eugenia, G.G.: *Data Mining in Medical and Biological Research*. In-Tech Publisher (2008)
12. An, A., Huang, Y., Huang, X., Cercone, N.: Feature Selection with Rough Sets for Web Page Classification. In: Peters, J.F., Skowron, A., Dubois, D., Grzymała-Busse, J.W., Inuiguchi, M., Polkowski, L. (eds.) *Transactions on Rough Sets II*. LNCS, vol. 3135, pp. 1–13. Springer, Heidelberg (2004)
13. Raymer, M.L., et al.: Dimensionality reduction using genetic algorithms. *IEEE Transactions on Evolutionary Computation* 4(2), 164–171 (2000)
14. Das, A.K., Sil, J.: An Efficient Classifier Design Integrating Rough Set and Set Oriented Database Operations. *Applied Soft Computing Journal* (2010), <http://dx.doi.org/10.1016/j.asoc.2010.08.008>
15. Dunn, J.C.: Well separated clusters and optimal fuzzy partitions. *J. Cybern.* 4, 95–104 (1974)
16. Davies, D.L., Bouldin, D.W.: Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1(2), 95–104 (1979)
17. Chou, C.H., Su, M.C., Lai, E.: A new cluster validity measure and its application to image compression. *Pattern Analysis and Applications* 7, 205–220 (2004)
18. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: Cluster validity methods: part II. *SIGMOD Rec.* 31(3), 19–27 (2002)
19. Ou, S.H.: *Rice Diseases*. Kew Surrey. Commonwealth Mycological Institute, Cambrian News(Aberystwyth) Ltd., England, Great Britain (1985)
20. International Rice Research Institute, Philippines, <http://www.irri.org>
21. George, G.D.: Biological and chemical control of rice blast disease (*Pyricularia oryzae*) in Northern Greece. *Cahiers Options Méditerranéennes* 15(3), 61–68
22. Otsu, N.: A Threshold Selection Method from Gray Level Histograms. *IEEE Transaction on Systems, Man and Cybernetics* 9, 62–66 (1979)
23. Hu, M.K.: Visual pattern recognition by moment invariants. *IEEE Transactions on Information Theory* 8, 179–187 (1962)
24. WEKA, *Machine Learning Software* (2011), <http://www.cs.waikato.ac.nz/~ml/>

A Novel GA-SVM Based Multistage Approach for Recognition of Handwritten *Bangla* Compound Characters

Nibaran Das*, Kallol Acharya, Ram Sarkar,
Subhadip Basu, Mahantapas Kundu, and Mita Nasipuri

Computer Sc. & Engg. Dept., Jadavpur University, Kolkata-700032, India
nibaran@ieee.org

Abstract. In the present work, a novel Genetic Algorithm (GA) and Support Vector Machine (SVM) based multistage recognition strategy has been developed to recognize handwritten *Bangla* Compound characters. The developed algorithm identifies optimal local discriminating regions in the second pass of the multistage approach, within each group of pattern classes identified by the first pass classifier. The developed technique has been used to evaluate handwritten *Bangla* Compound characters having 8254 numbers of samples of 171 character classes. These 171 classes of characters are eventually distributed among 199 pattern classes, where some character classes share multiple pattern shapes. Employing the GA-SVM powered region optimization in the second pass, we have obtained an accuracy of 78.93% on 171 character classes, which is a clear 2.83% improvement over the result achieved by the corresponding single pass approach.

1 Introduction

In any complex character recognition problem, multiple classifiers are often used to decide on the label of an unknown test pattern. In such systems, classifiers are either used concurrently or in sequential stages. In a concurrent *classifier combination approach* [1], the weakness of one classifier can be complemented by the strength of another classifier. In contrast, a *multi-stage* approach [2], is preferred by the research communities due to its simple and hierarchical approach. For example, in a typical two-pass classification scheme, the first-stage classifier produces a coarse classification decision on the basis of some global features extracted from the unknown pattern. The next-stage classifier(s) then refines the coarse classification decision to produce the final classification decision on the basis of some local features extracted from the character pattern. Actions of the two classifiers, which work in sequence, thus constitute two passes of the classification process. While the design of the first-pass classifier is relatively simple but in the second-pass challenge exists in selection of optimal local features from discriminating sub-images. This motivates us to develop a novel Genetic Algorithm (GA)–Support Vector Machine (SVM) based multistage strategy for recognition of handwritten *Bangla* Compound characters.

* Corresponding author.

2 The Present Work

In the present work, we have developed multistage recognition strategy using GA-SVM combination for recognition of handwritten *Bangla* Compound characters. *Bangla*, the second most popular script in India is enriched with more than 260 Compound characters apart from 50 Basic characters and 10 Numerals. A Compound character is a complex shaped character, which consists of two or more Basic characters, pronounced simultaneously in words of *Bangla* language. The shapes of compound characters are very complex and some compound characters resemble pair-wise so closely that the only small sign of differences such as short or long straight lines, circular curves etc. left between them. Therefore, it is often difficult to identify those characters without analyzing the context, especially for handwritten documents. Moreover the shapes and number of compound characters varies with time. From a recent survey conducted by the same author, it has been revealed that the count of popularly used compound characters is about 269. But for the present work, we have considered only 171 compound characters after discarding the characters with *Reph* and *Ya-phala*. Also, due to having more than one shape for some compound characters, the number of pattern classes becomes 199. Therefore, apart from the large number of pattern classes, here we have to deal with those shape similarities.

Some research have already been done on recognition of handwritten Basic characters, and numerals of *Bangla* script[2-4]. Despite such efforts, the OCR of *Bangla* script remains incomplete without detailed research on compound characters. There are three earlier instances of works on handwritten *Bangla* Compound characters[5-7] in the literature. In[5], Pal et al. used Modified Quadratic Discriminant Function (MQDF) with the directional information obtained from the arc tangent of the gradient to recognize 138 class compound characters of 20543 samples. In one of our earlier works,[6] a technique for recognition of handwritten *Bangla* Compound characters was proposed with a discussion on the potential problems of *Bangla* Compound character recognition. The methodology was used for top 55 characters according to their frequency using quad tree based longest run feature. In another work[7], the 55 compound characters combined with 50 basic characters were used for recognition. Shadow and quad tree based longest run features are used there with Multi-Layer Perceptron (MLP) and SVM. From the above discussion, it is evident that *Bangla* Compound character recognition is a difficult, yet essential and under studied problem in OCR research.

From our earlier experience on development of handwritten OCR researches[8], we have found that the use of multistage/hierarchical approach instead of single stage approach ameliorates the recognition accuracy, especially when the number of classes is large. This is because, usage of designed-for-all global features is found to be incapable in estimating reasonable class boundaries in the feature space for large number of classes.

In the first stage of the present work, all the 199 characters are classified using SVM classifier based on some global feature descriptors along with the local features covering the entire image. From the confusion matrix prepared from the classification result we have found that there exist some groups of character patterns for which within-group mutual misclassification rate is high. We have attempted to utilize this potentially useful information and developed a pre-classification group formation

algorithm to identify potential clusters of mutually misclassifying patterns from the said confusion matrix.

In the finer classification stage, for each multiclass group of pattern classes, separate classifiers are custom designed with prudent choice of local features. In this stage, the selection of optimum local features is done heuristically from the region(s) where the patterns of different classes of the same group differ significantly from each other. A GA[9]based optimization strategy is employed in this work for this purpose.

After finer classification of different groups of patterns, the total number of classified pattern classes still remains 199, whereas the actual number of compound character classes is 171. We then use a mapping technique to finally get back the actual class labels and evaluate the recognition accuracy. Fig.1shows a schematic structure of the designed framework, while the following subsection describes the different global and local features.

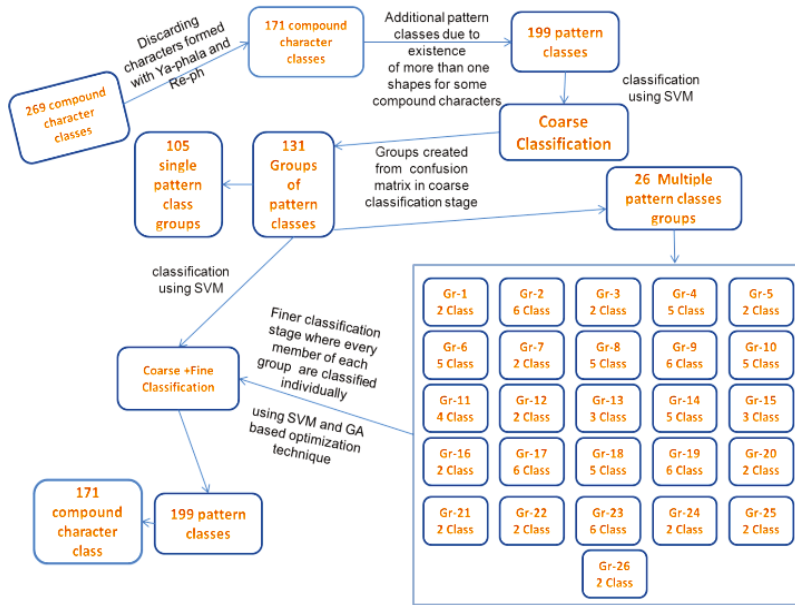


Fig. 1. Schematic structure of the present work

2.1 Design of the Feature Set

For classification of character patterns we have used different structural/topological features extracted either from the entire pattern image (global features) or from some local regions of the image (local features). We have used modified shadow features[7], octant-centroid features[7], Quad-tree based longest run features[7, 10] and different topological attributes of the character patterns like loop count[10], diagonal and angular distances as global features. Longest run features are also used as local features in different groups of characters in addition to the global features. A brief introduction of different distance and diagonal features are described below.

Distance. For computing these features, some fiducial points on the character patterns are identified and the distance between a pair of such points is considered as a topological information about the character pattern. Description of various distance features, considered in the present work are given below. All the distance values are normalized by dividing each of them either by the length of diagonal of the bounding rectangle (for feature#1-8) or by the horizontal width of the bounding rectangle (for feature #9-12).

Angle. Angle subtended by a line joining two fiducial points on the character patterns with respect to the horizontal direction is another type of topological feature considered in the present work. Here the angle subtended by the lines joining the fiducial points of distance features #3 and #4 with the horizontal direction are taken as two angle features (Fig. 2(h)).

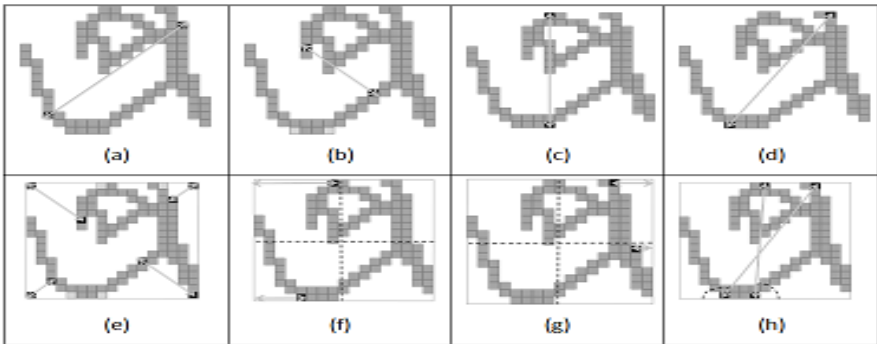


Fig. 2. Illustration of different distances and angle features (a)Features #1 (b)Features #2 (c)Features #3 (d)Features #4 (e)Features #5-8 (f)Features #9-10 (g) Features #11-12 (h) Angle subtended by distance feature 3 and 4 with the horizontal line

Design of the multi stage character recognizer. We have evaluated the above multi stage approach on the handwritten *Bangla* Compound character database[11] entitled as CMATERdb 3.1.3 which was prepared in CMATER Laboratory, Jadavpur University, Kolkata, India. The database consists of 33282 training and 8254 test samples. It is noteworthy to mention that before evaluation all the pattern classes considered in our experiment are normalized to 96X96 pixels after binarization.

First stage classification. It is already been mentioned that our dataset is divided into training and test sets. For formation of groups, first we have generated a validation set consisting of randomly chosen $1/3^{\text{rd}}$ data samples from each training class. Then, the rest of the samples of training classes are used to train with SVM classifier using all 139 features described above. The trained classifier is used to obtain the recognition accuracy on validation set. The confusion matrix generated after adding the three confusion matrices over the different validation set is used to identify different groups of mutually misclassifying patterns. An automatic grouping algorithm has been developed to select the number of groups and their members from the large confusion

matrix (of order 199x199). From the algorithm we have obtained 26 pattern groups having at least two mutually misclassifying pattern classes. 94 pattern classes belonging to the above 26 groups are shown in Fig. 3. The remaining 105 pattern classes are considered as singleton groups.

After formation of groups, all the pattern classes of each multiclass pattern groups are labeled with a single group identity and SVM classifier is again trained for all the multiclass pattern groups along with single class pattern groups using all the 139 local and global features. The main objective of it is to increase the recognition accuracy for each group. This is highly desirable since if an unknown pattern class is wrongly classified into a group other than the true one in the first stage, nothing could be done to refine this decision in the second stage.

Table 1. Description of different distance based features

Feature #	Chosen fiducial points
1	Two farthest character pixels along the diagonal from top-left corner to the bottom-right corner of the bounding rectangle(Fig.2(a)).
2	Two farthest character pixels along the diagonal from top right corner to bottom left corner of the bounding rectangle(Fig 2.(b)).
3	Top left and bottom right character pixels(Fig. 2(c)).
4	Top right and bottom left character pixels(Fig. 2(d)).
5-8	The corner points of the bounding rectangle and the corresponding diagonally nearest character pixels. (Fig. 2(e)).
9-10	Left edge of the bounding rectangle and horizontally farthest rightmost character pixels in quadrants 2 and 3(Fig. 2(f)).
11-12	Right edge of the bounding rectangle and horizontally farthest rightmost character pixels in quadrants 1 and 4(Fig. 2(g)).

Second stage Classifiers for individual groups. For discrimination between similar looking pattern shapes, human cognition system generally tries to locate the region(s) of dissimilarity among the pattern shapes in each group and then focuses on those regions to recognize the pattern shapes. In the first stage classification we have considered local features from all local regions covering the entire character image so that the recognition accuracy of each group remains high. In the second stage, our objective will be to find out the set of local regions for each group which provides discriminatory information about the pattern classes in that group setting aside the other local regions which might lead to ambiguity. But identification of the proper set of regions needs thorough experimentation because it requires exploring huge search space if we check all the combinations of local regions. To overcome the problem we have applied GA[9], a heuristic iterative searching methodology which is free from the problem of sticking at local minima. GA is also used to find out the optimal set of sub-regions on the character images for each group of character classes. To implement GA, we have used 64 longest run features extracting from 16 regions generated from quad tree of depth two, partitioning along with global features described in section 2.1. During implementation of GA, various combinations of local regions are represented as chromosomes where each gene corresponds to a particular

local region is represented by a single bit of a 16 bit chromosome string; the n^{th} bit of the chromosome is set to 1 if the n^{th} local region is considered for calculation. For each generation, we have considered 50 numbers of chromosomes among which 40% are selected rank wise in descending order and rest 60% are selected by roulette wheel for next generation keeping chromosome number fixed for every generation. From the generated chromosomes 80% chromosomes are selected pair wise randomly for crossover. For the present work, 20 randomly selected chromosomes are used for single bit mutation. The bit position is also selected randomly for the chromosome. In this work, the stopping criterion is reached either after *significant* generations (maximum 50 generations) have passed or the average fitness value of the current population is greater than or equal to 99% of the maximum fitness value obtained. The classification decision produced by the chromosome with the highest fitness value is taken as final.

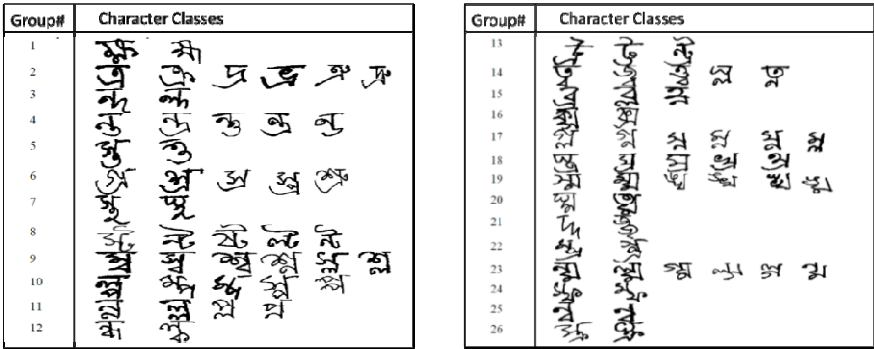


Fig. 3. Images samples of different groups are shown in a single row with their group number

3 Experimental Results

As mentioned before, the pattern classifiers used for this work are all designed with SVM. For this purpose, we have used LIBSVM[12] with Radial Basis Function (RBF) kernel. At the first stage which is also called as single pass approach, we have obtained recognition accuracy of 75.83% using all the 139 features for the 199 pattern class on test set. It has already mentioned that 26 multiclass groups have been obtained after first pass. The pattern classes of a multiclass group are labeled as single label with their group identity. In this way number of classes reduced to 131 from 199, initially considered for coarse classification. The obtained recognition accuracy is 84.55% for this 131 class. Recognition accuracy of different groups on test set using SVM based classifiers with all the sub-regions and optimally selected regions based on GA are presented in Table 2. It can be observed from the Table 2 that GA based region selection strategy always provides higher recognition accuracy than the recognition accuracy achieved by considering the features extracted from all regions. Table 2 also demonstrates the merit of GA for optimally selecting local regions over

the all local regions both in terms of recognition accuracy as well as the number of features. Here true classification is done after combining classification result of 102 classes with the group classification. Thus we get recognition of 78.93%. It is worthy to mention that in the multistage approach proper coarse classification is not possible at the first stage due to the existence of multiclass in the same group. The recognition performance of the multistage classification is compared with the coarse classification designed for the same. We have obtained recognition accuracy of 76.10% for the coarse classification of 171 character classes. Thus an improvement of 2.83% is observed after implementation of multistage approach. As direct comparison of the obtained result is not possible due to lack of a globally accessible database for handwritten *Bangla* Compound characters, here we have designed Table 3 for citing other sort of works on compound character recognition.

Table 2. Comparison of the recognition rates obtained with features extracted from all local regions and with features from optimally selected regions

Group #	Recognition rate (%)			Number of regions selected	Group #	Recognition rate (%)			Number of regions selected by GA
	Features from all local regions	Optimally selected regions by GA	Increments			Features from all local regions	Optimally selected regions by GA	Increments	
1	92.05	94.318	2.268	7	14	76.1	87.61	11.51	9
2	71.62	84.08	12.46	5	15	79.05	97.29	18.24	5
3	85.41	95.83	10.42	5	16	75	90.38	15.38	5
4	63.41	71.54	8.13	10	17	74.14	84.79	10.65	3
5	96.93	100	3.07	4	18	69.54	83.63	14.09	5
6	81.28	88.17	6.89	9	19	70.83	80	9.17	2
7	82.14	96.42	14.28	2	20	74.66	93.33	18.67	5
8	67.41	83.031	15.62	6	21	97	98	1	6
9	58.84	79.83	20.99	4	22	73.23	91.54	18.31	7
10	72.67	81.39	8.72	6	23	71.36	84.23	12.87	4
11	67.34	84.69	17.35	7	24	78.78	86.36	7.58	3
12	92.53	97.87	5.34	5	25	91.39	94.62	3.23	3
13	86.66	94.07	7.41	5	26	80.769	86.53	5.76	7

Table 3. Comparative overview of handwritten *Bangla* Compound character

The work reference	Number of samples taken for forming the database		Number of Class Samples	Classifier/ Classification scheme	Recognition accuracy on the test set
	Training	Testing			
U. Pal[5]	5 fold cross validation of 20543 class samples		138	Modified Quadratic Discriminant Function (MQDF)	85.90%
N. Das[6]	3 fold cross validation of 9765 class samples		55 reduced to 43	Multi-layer perceptron	84.67%
N.Das[7]	3 fold cross validation of 19765 class samples		93 class(50 Basic+ 43 Compound)	Support Vector Machine	80.51%
Present work	33282	8254	171	Support Vector Machine	78.93

4 Conclusion

In the current work, we have developed a novel GA-SVM based multistage approach for handwritten *Bangla* Compound characters having large number of classes. The approach is better than the general two pass approach [13] due to identification of optimal local discriminating regions within a group discarding the regions containing ambiguous information. It is designed to boost both recognition accuracy as well as speed since the applied features have been extracted only from those local regions which play a significant role in identifying the pattern classes. Apart from this, it is the first work on handwritten *Bangla* Compound characters available nowadays. Introduction of stronger feature set in global and local regions may be included in future to improve the recognition accuracy.

Acknowledgements. Authors are thankful to the CMATER, SRUVM, and PURSE project of Computer Science & Engineering Department, Jadavpur University, for providing infrastructure facilities during progress of the work.

References

1. Ho, T.K.: Multiple Classifier Combination: Lessons and Next Steps. In: Hybrid Methods In Pattern Recognition, pp. 171–198. World Scientific (2002)
2. Bhattacharya, U., Chaudhuri, B.B.: Handwritten Numeral Databases of Indian Scripts and Multistage Recognition of Mixed Numerals. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 444–457 (2009)
3. Bhowmik, T., Ghanty, P., Roy, A., Parui, S.: SVM-based hierarchical architectures for handwritten Bangla character recognition. *International Journal on Document Analysis and Recognition* 12, 97–108 (2009)
4. Liu, C.-L., Suen, C.Y.: A new benchmark on the recognition of handwritten Bangla and Farsi numeral characters. *Pattern Recognition* 42, 3287–3295 (2009)
5. Pal, U., Wakabayashi, T., Kimura, F.: Handwritten Bangla Compound Character Recognition Using Gradient Feature. In: 10th International Conference on Information Technology (ICIT 2007), pp. 208–213 (2007)
6. Das, N., Basu, S., Sarkar, R., Kundu, M., Nasipuri, M., Basu, D.K.: Handwritten Bangla Compound character recognition: Potential challenges and probable solution. In: 4th Indian International Conference on Artificial Intelligence, pp. 1901–1913 (2009)
7. Das, N., Das, B., Sarkar, R., Basu, S., Kundu, M., Nasipuri, M.: Handwritten Bangla Basic and Compound character recognition using MLP and SVM classifier. *Journal Of Computing* 2 (2010)
8. Basu, S., Das, N., Sarkar, R., Kundu, M., Nasipuri, M., Basu, D.K.: A hierarchical approach to recognition of handwritten Bangla characters. *Pattern Recognition* 42, 1467–1484 (2009)
9. Srinivas, M., Patnaik, L.M.: Genetic algorithms: a survey. *Computer* 27, 17–26 (1994)
10. Das, N., Basu, S., Sarkar, R., Kundu, M., Nasipuri, M., Basu, D.K.: An Improved Feature Descriptor for Recognition of Handwritten Bangla Alphabet. In: International Conference on Signal and Image Processing, pp. 451–454. Excel India Publishers (2009)
11. <http://code.google.com/p/cmaterdb/> (accessed August 21, 2011)
12. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2, 1–27 (2011)
13. Bhattacharya, U., Shridhar, M., Parui, S.K.: On Recognition of Handwritten Bangla Characters. In: Kalra, P.K., Peleg, S. (eds.) *ICVGIP 2006*. LNCS, vol. 4338, pp. 817–828. Springer, Heidelberg (2006)

Cascaded Correlation Neural Network Based Microcalcification Detection in Mammographic Images

J. Dheeba¹ and S. Tamil Selvi²

¹ Department of Computer Science and Engineering,
Noorul Islam University, Kumaracoil, TN
deeps_3u4@yahoo.com

² Department of Electronics and Communication Engineering
National Engineering College, Kovilpatti, TN
tamilgopal2004@yahoo.co.in

Abstract. This paper presents a novel approach for classification of microcalcification (MC) clusters in mammograms. This cluster is the significant indication of breast cancer in women at the early stage. Diagnosis of these clusters at the early stage is a very difficult task as the cancerous tumors are embedded in normal breast tissue structures. This paper proposes an artificial intelligent neural network algorithm - Cascaded Correlation Neural Network (CCNN) - for detection of tumors in mammograms. CCNN has a distinct feature that it does not use a predefined set of hidden units, instead the hidden units gets added up one by one until the error is minimized. By exploiting this distinct feature of the CCNN, a computerized detection algorithm is developed that are not only accurate but also computationally efficient for microcalcification detection in mammograms. Prior to MC detection texture features from the Region of Interest (ROI) of the mammographic Image is extracted using gabor features. Then CCNN classifier is used to determine whether the input data is normal/benign/malignant. The performance of this scheme is evaluated using a database of 322 mammograms from MIAS database and real time clinical mammograms. The result shows that the proposed CCNN algorithm has good performance.

Keywords: Computer Aided Diagnosis, Microcalcification, Mammograms, Artificial Intelligence, Cascaded Correlation Neural Network, Texture features.

1 Introduction

Breast cancer is second frequently diagnosed cancer among women, especially in developed countries. In western countries about 53%-92% of the population has this disease. In a Phillipine study [1] a mammogram screening was done to 151,198 women. Out of that 3479 women had this disease and were referred for diagnosis. Though breast cancer leads to death, early detection of breast cancer can increase the survival rate. The current diagnostic method for early detection of breast cancer is mammography. Mammographies are low dose X-ray projections of the breast, and it is the best method for detecting cancer at the early stage.

Microcalcifications (MC) are quiet tiny bits of calcium, and may show up in clusters or in patterns and are associated with extra cell activity in breast tissue. Usually the extra cell growth is not cancerous, but sometimes tight clusters of microcalcification can indicate early breast cancer. Scattered microcalcifications are usually a sign of benign breast cancer. 80% of the MC is benign. MC in the breast shows up as white speckles on breast X-rays. The calcifications are small; usually varying from 100 micrometer to 300 micrometer, but in reality may be as large as 2mm. Though it is very difficult to detect the calcifications as such, when more than 10 calcifications are clustered together, it becomes possible to diagnose malignant disease. But the survival depends on how early the cancer is detected. So, any MC formation should be detected at the benign stage. Hence, a Computer Aided Diagnosis (CAD) system is used to detect MC clusters [16, 17].

Many different algorithms have been proposed for automatic detection of breast cancer in mammograms. Features extracted from mammograms can be used for detection of cancers [2]. Studies reports that features are extracted from the individual MCs [3, 15] or from the ROI which contain MC clusters [4].

Yu and Ling [7] has proposed a CAD system by employing a mixed feature set. Using the mixed feature set of 31 features the true microcalcification pixel is identified. The discriminatory power of these features is analyzed using general regression neural networks via sequential forward and sequential backward selection methods. Netsch and Heinz-Otto [8], uses the Laplacian scale-space representation of the mammogram. First, possible locations of microcalcifications are identified as local maxima in the filtered image on a range of scales. For each finding, the size and local contrast is estimated, based on the Laplacian response denoted as the scale-space signature. A finding is marked as a microcalcification if the estimated contrast is larger than a predefined threshold which depends on the size of the finding.

Berman Sahiner et al. used a Convolution Neural Network (CNN) classifier to classifier the masses and the normal breast tissue [9]. First, the Region of Interest (ROI) of the image is taken and it was subjected to averaging and subsampling. Second, gray level difference statistics (GLDS) and spatial gray level dependence (SGLD) features were computed from different subregions. The computed features were given as input to the CNN classifier.

Cascio.D [10] developed an automatic CAD scheme for mammographic interpretations. The scheme makes use of the Artificial Neural Network to classify mass lesions using the geometric information and shape parameters as input to the classifier. Jong and Hyyun [11] proposed a three layer Backpropagation Neural Network (BPNN) for automatic detection of microcalcificaiton clusters. Texture features are extracted to classify the ROI containing clustered MC and ROI containing normal tissues.

These computerized methods discussed in the previous paragraphs did not analyze the texture component of the image in detail. Gabor features [13] are used in several image analysis applications including texture classification and proved to be efficient.

The major objective of this paper is to take multiple texture features from the original image to discriminate between microcalcification and the normal tissue in the breast. As a first stage, the original image is preprocessed and Region of Interest (ROI) is taken and features are extracted from the ROI image using gabor features. In the second stage, the extracted features are compared by means of their ability in detecting microcalcification clusters using CCNN. We use mammograms from the

Mammographic Image Analysis Society (MIAS) database which contain 322 mammograms [12] and real time clinical mammograms.

2 Proposed Methodology

The block diagram of the proposed MC detection methodology is shown in Fig 1. The Digitized Mammogram Image is preprocessed and the goal of preprocessing is to simplify recognition of MC without throwing away any important information. As a preprocessing step the breast area is separated from the background image. The breast area is chosen as ROI for the next stage of processing. This saves the processing time and also the memory space.

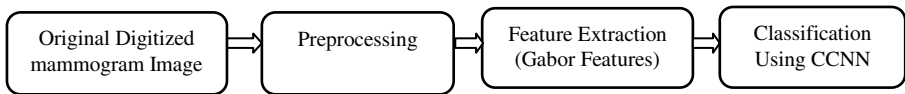


Fig. 1. Block Diagram of the Proposed System

3 Feature Extraction

In image processing the texture of a region describes the pattern of spatial variation of gray tones (or in the different color bands in a color image) in a neighborhood that is small compared to the region. The Gabor wavelet was first introduced by David Gabor in 1946. The most important properties are related to invariance, illumination, rotation, scale, and translation. These properties are especially useful in feature extraction, where Gabor filters have succeeded in diverse applications, like texture analysis.

Gabor filters has been used by many authors to extract the local texture features [5] [6]. The input image $I(x, y)$ is convolved with the gabor function $g(x, y)$.

The 2-D gabor function $g(x, y)$ is given by,

$$g(x, y, \omega, \theta, \sigma_x, \sigma_y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp \left[\frac{-1}{2} \left(\left(\frac{x}{\sigma_x} \right)^2 + \left(\frac{y}{\sigma_y} \right)^2 \right) + j\omega(x\cos\theta + y\sin\theta) \right]$$

Where σ is the spatial spread, ω is the frequency, θ is the orientation. $g(x, y, \omega, \theta, \sigma_x, \sigma_y)$ forms the mother gabor wavelet and then this self-similar filter dictionary can be obtained by appropriate dilations and rotations through the generating function:

$$g_{mn}(x, y) = a^{-m} G(x', y'), a > 1, m, n = \text{integer}$$

$$x' = a^{-m} (x \cos \theta + y \sin \theta), \text{ and } y' = a^{-m} (-x \sin \theta + y \cos \theta),$$

where, $\theta = n\pi / K$ and K is the total number of orientations. The scale factor a^{-m} is meant to ensure that the energy is independent of m . In order to eliminate sensitivity of the filter response to absolute intensity values, the real components of the gabor filters are biased by adding a constant to make them zero mean.

4 Proposed CCNN Based MC Classification

In order to find the potential microcalcification pixels based on the above mentioned features, a proper classification method must be used. In our study, the classifier chosen is a Cascaded Correlation neural network [14]. The main drawback in artificial neural network is the rate of convergence and the manual fixation of network architecture (hidden units) throughout training. These problems are addressed by cascaded correlation neural network. Moreover it uses simple training rules since only one layer of weights is being trained at a time.

A cascaded correlation network possesses input units, hidden units and output units that are connected directly to output unit with adjustable weighted connections. Connection from input unit to the hidden unit are trained when the hidden unit is added to the net and then they are frozen. The connection from the hidden unit to the output unit is also adjustable.

The training data set for the proposed system is chosen from the 115 ROI's with a cluster of microcalcifications in the center. If one microcalcification pixel is picked out from one ROI, a corresponding normal pixel is chosen randomly from the same ROI. This makes the number of microcalcification pixels equal to the number of normal pixels in the training data set. After training, the neural network is used to classify the 207 full mammograms in the database.

Gabor features are given as input to the network for training. The desired output from the network is whether the MC is present or not present. Hidden layers and neurons are chosen automatically. During the training session of the network a pair of patterns is presented, the input pattern (gabor features) and the target or the desired pattern (abnormal or normal). At the output layer, the difference between the actual and target outputs yields an error signal. This error signal depends on the values of the weights of the neurons in each layer. This error is minimized, and during this process new values for the weights are obtained.

Steps involved in forming the cascaded network are,

1. The network as shown in Fig.2 starts with input (gabor features) and output unit as whether the tissue is normal/benign/malignant.
2. This simple network is trained and the error is calculated using the equation below

$$E_j(p) = y_j(p) - d_j(p) \quad j = 1 \quad (1)$$

Where p is the total number of training patterns, $E_j(p)$ is the residual error is for the output unit, $y_j(p)$ is the actual output, $d_j(p)$ is the desired output.

3. Then the hidden units are added.
4. A temporary unit (z) is added and it is connected to the input unit.
 - a. The weights from the input unit to the temporary unit are adjusted.
 - b. The error is calculated using equation 1, then $E_j(p)$ is multiplied by the derivative of the output unit activation function.

$$z - av = \frac{1}{p} \sum_{p=1}^P z(p) \quad (2)$$

Where $z - av$ is the average activation of temporary unit. Sigmoid activation function is used.

- c. After the training process is completed, the weights are frozen and the temporary unit becomes the permanent hidden unit.
5. Now, a new hidden unit is connected to the output unit.
 - a. The weights from the new hidden unit to the output unit are adjusted,
 - b. Now, the connections to the output unit are trained.
 - c. The correlation between the output units and the temporary hidden units is expressed as,

$$S = \sum_j \sum_p (z(p) - (z - av))(E_j(p) - (E - av_j))$$

6. The whole process is continued until the error reaches an acceptable level.

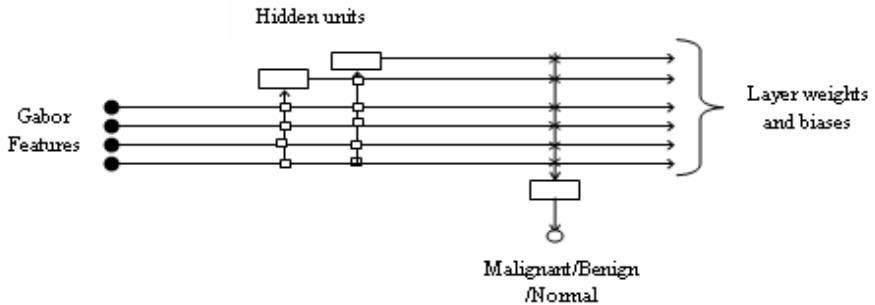


Fig. 2. CCNN Architecture

5 Experimental Results

The performance of the proposed methodology was tested on digitized mammograms from the MIAS database provided by the Mammographic Image Analysis Society (MIAS) in the UK [12]. The database contains left and right breast images of 161 patients. Its quantity consists of 322 images, which belongs to three types such as Normal, benign and malignant. The database has been reduced to 200 micron pixel edge, so that all images are 1024×1024 . There are 208 normal, 63 benign and 51 malignant (abnormal) images. The database images have four different kinds of abnormalities namely: architectural distortions, stellate lesions, Circumscribed masses and calcifications. The CCNN algorithm classifies the input image into suspicious and non-suspicious regions. For the classification experiments, the training dataset contain a total of 2160 gabor patterns obtained from the mammograms. These patterns contains pixels including true individual microcalcification clusters, circumscribed masses, ill

defined masses and also pixels indicating normal tissues that includes blood vessels and dense breast tissues. The classification results are shown in Fig. 3. The detection results shows a malignant tumor mammogram pattern, which is classified by the proposed method. The tumor affected portion is indicated by blue circle.

The main aim of the proposed system was that no case of malignancy-indicating microcalcification should escape radiological analysis. We therefore started from two basic assumptions: (i) the microcalcifications have an aspect that differentiates them from the other elements of the breast because of their different X-ray opacity; and (ii) since we are looking for microcalcifications that are in an incipient stage, they involve a very small proportion of the total area of the breast because they otherwise would be clearly visible to any radiologist and there would consequently be no point in using our system.

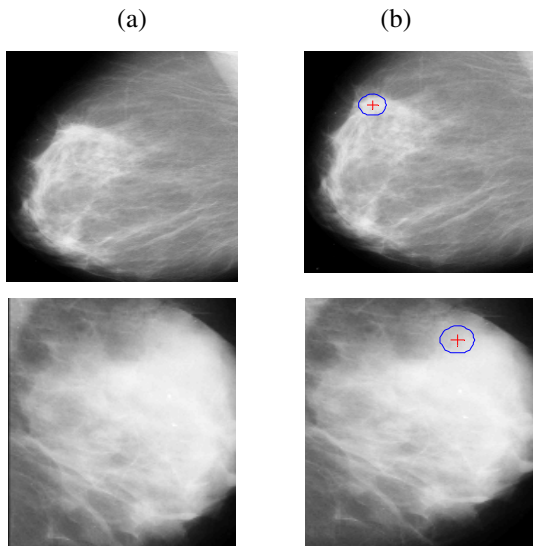


Fig. 3. (a) MIAS Mammogram Image (ROI) (b) CCNN Classified Tumor Patterns

Breast cancer screening is generally based on two-view mammography in which mediolateral oblique (MLO) and a cranio caudal (CC) projections are obtained from both breasts. When reading mammograms, radiologists combine information from all available views. In this study, we focus on development of a CAD system for the detection of tumors with the help of different views (MLO and CC views) using mammographic images collected from mammogram screening centers. If a suspicious region in one view has certain features in common with a suspicious region in the other view, there is a higher probability that the region is a true microcalcification.

Real time clinical mammograms were collected from 54 patients and all these patients have agreed to have their mammograms to be used in research studies. For each patient 4 mammograms were taken in two different views, one is the Craniocaudal (CC) and the other is the Mediolateral Oblique (MLO) view. The two projections of

each breast (right and left) were taken for every case. For this study a total of 216 mammograms were taken, all the mammograms were digitized to a resolution of 290 x 290 Dots per Inch (DPI) which produces 24 bits/pixel. Each digitized mammograms was incorporated into a 2020 x 2708 pixel image (5.47 Mpixels).

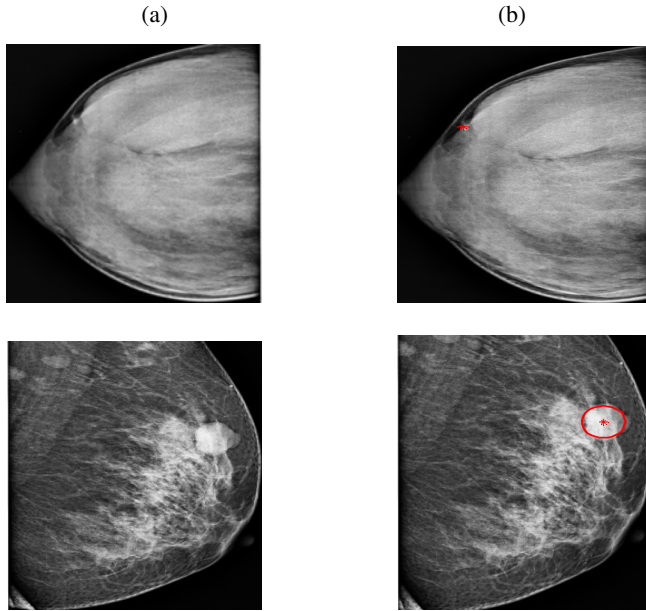


Fig. 4. (a) Real time Clinical Mammogram Image (ROI) (b) CCNN Classified Microcalcification

Fig 4 (a) shows the ROI image of a digital mammogram in and Fig 4(b) shows the CCNN classified microcalcification marked with red circles. The detected regions in a mammogram corresponding to tumor tissues have different texture patterns and gray levels than the normal ones and by employing CCNN classifier it is possible to classify these regions. The computerized scheme based on the CCNN classifier with gabor features for distinguishing among different types of abnormalities in the test set, yields a classification accuracy of 89.32% and an accuracy of 85.17% for real time clinical images.

6 Conclusion

The novel approach presented in this paper demonstrated that the CCNN classifier produces an improvement in classification accuracy to the problem of computer-aided analysis of digital mammograms for breast cancer detection. The algorithm developed here classifies mammograms into normal, benign and malignant. First, the ROI of the image is chosen then gabor features are extracted and classified using Cascaded Correlation neural networks. Using the mammographic data from the Mammographic Image Analysis Society (MIAS) database classification accuracy of 89.32% and a score of 85.17% for real time clinical images is achieved using the proposed approach.

References

1. Pisani, et al.: Outcome of screening by Clinical Examination of the Breast in a Trial in the Phillipines. *Int. J. Cancer* (2006)
2. Shen, L., Rangaan, R.M., Desautels, J.E.L.: Application of shape analysis to mammographic classifications. *IEEE Trans. Medical Imaging* 13(2), 263–274 (1994)
3. Lee, S.K., Chung, P., Chang, C.L., Lo, C.S., Lee, T., Hsu, G.C., Wang, C.: Classification of Clustered Microcalcifications using shape cognitron neural network. *Neural Networks* 16(1), 121–132 (2003)
4. Dhawan, A.P., Chitre, Kaiser-Bonasso, C., Moskoitz, M.: Analysis of mammographic microcalcification using gray-level image structure features. *IEEE Trans. Medical Imaging* 15(3), 11–150 (2005)
5. Tan, T.N.: Texture edge detection by modeling visual cortical channels. *Pattern Recognition* 28(9), 1283–1298 (1995)
6. Turner, M.R.: Texture discrimination by Gabor functions. *Biol. Cybern.* 55, 71–82 (1986)
7. Yu, Ling: A CAD System for the Automatic Detection of Clustered Microcalcifications in Digitized Mammogram Films. *IEEE Transactions on Medical Imaging* 19(2), 115–126 (2000)
8. Netsch, Heinz-Otto: Scale-Space Signatures for the Detection of Clustered Microcalcifications in Digital Mammograms. *IEEE Transactions on Medical Imaging* 18(9), 774–786 (1999)
9. Sahiner, B., et al.: Classificaiton of Mass and Normal Breast Tissue: A convolution Neural Network classifier with spatial domain and Texture Images. *IEEE Trans. On Medical Imaging* 15(5), 598–609 (1996)
10. Cascio, D., et al.: Mammogram Segmentation by Contour Searching and Mass Lesions Classification with Neural Network. *IEEE Trans. On Nuclear Science* 53(5), 2827–2833 (2006)
11. Kim, J.K., Park, H.W.: Statistical Textural Features for Detection of Microcalcifications in Digitized Mammograms. *IEEE Trans. on Medical Imaging* 18(3), 231–238 (1999)
12. Suckling, J., Parker, J., et al.: The mammographic images analysis society digital mammogram database. In: *Proc. 2nd Int. Workshop Digital Mammography*, ork, U.K, pp. 375–378 (July 1994)
13. Manjunath, B.S., Chellappa, R.: A Unified Approach to Boundary Detection. *IEEE Trans. Neural Networks* 4(1), 96–108 (1993)
14. Fahlman, S.E., Lebiere, C.: The Cascade-Correlation Learning Architecture. In: Touretzky, D. (ed.) *Neural Information Processing Systems*, pp. 524–532. Morgan Kaufmann Publishers, Inc., Denver (1990)
15. Wong, A., Scharcanski, J.: Phase-Adaptive Superresolution of Mammographic Images Using Complex Wavelets. *IEEE Transactions on Image Processing* 18(5), 1140–1146 (2009)
16. Tsui, P.-H., Liao, Y.-Y., Chang, C.-C., Kuo, W.-H., Chang, K.-J., Yeh, C.-K.: Classification of Benign and Malignant Breast Tumors by 2-D Analysis Based on Contour Description and Scatterer Characterization. *IEEE Transaction on Medical Imaging* 29(2), 513–522 (2010)
17. Mencattini, A., Salmeri, M., Rabottino, G., Salicone, S.: Metrological Characterization of a CADx System for the Classification of Breast Masses in Mammograms. *IEEE Transactions on Instrumentation and Measurement* 59(11), 2792–2799 (2010)

TertProt: A Protein Fold Recognition Method Using Protein Secondary Structure Program

D.S.V.G.K. Kaladhar

Department of Bioinformatics, GIS, GITAM University,
Visakhapatnam, India
dr.dowluru@gmail.com

Abstract. TertProt is a protein secondary structure program written in ANSI C language implemented using Chou-Fasman conformational parameters with tertiary structure prediction methodology. The secondary structure is a key element for the analysis and modelling of protein structure. The TertProt method for modeling of a 3D structure of a protein is done using Colloc'h triangle has been designed in the present research. The structures has been compared with Oxytocin, HCMV protease, D, L altering peptide and Alzheimer beta amyloid peptides.

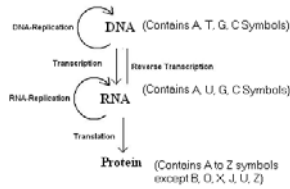
Keywords: Chou-Fasman, Colloc'h triangle, C program.

1 Introduction

Every living organism share many common attributes to tolerate, defend, mutate and manipulate the molecules by cellular process due to surrounding factors. Definition of secondary structure is essential, for a successful analysis of the relation between amino acid sequence and protein structure [1]. The ultimate goal of structural studies of proteins is to gain insight into protein three-dimensional structure at high-resolution level. This can often be accomplished by the application of techniques such as X-ray crystallography or multidimensional nuclear magnetic resonance (NMR) [2]. The recent advances in protein folding are reviewed based on a classification of the approaches in comparative modeling, fold recognition, and database information [3].

1.1 The Cell

The cell is the basic unit of life and has provided great impetus to the progress of information as biomolecules [4]. Double-stranded DNA molecules containing alphabets such as A, T, G, C, are duplicated by the process called replication. The triplet-code words of genetic information encoded in DNA sequence are transcribed into codons of messenger RNA (mRNA) with alphabets A, U, G, C, which in turn are translated into an amino acid sequence of polypeptide chains called proteins containing alphabets from A to Z except B, O, X, J, U, Z.



1.2 Protein

Among biomolecules, proteins are the most common molecules in living organisms involved in important cellular processes. Proteins are the cornerstones of cell structure and the agents of biological function evolving signals from DNA [5]. Studying the structure of proteins would help in better understanding about process of life in a living system. The expression of protein is tightly regulated for normal functioning of a cell or organism. A polypeptide chain usually has one free terminal amino group (N-terminal) and a terminal carboxyl group (C-terminal) at the other end, though sometimes they are derivatized. Polypeptide chains formed by polymerization of α -amino acids in specific sequences provide the primary structure of proteins. The regular secondary structures, α helices and β sheets, are connected by coil or loop regions of various lengths and irregular shapes. The secondary structures are combined with specific geometric arrangement to form compact globular structure known as tertiary structure. Polypeptide chains, especially of regulatory proteins, often aggregate by specific interactions to form oligomeric structures. These oligomeric proteins are said to exhibit quaternary structure.

1.3 Protein Secondary Structure

A polypeptide (protein) can be thought of as a chain of flat peptide units for which each peptide unit is connected by the α -carbon of an amino acid [6]. The secondary folding of an amino acid chain into an actively stable structure is important in protein-ligand or protein-protein interaction studies. Two common examples are the α -helix and the β -pleated sheet. These shapes are reinforced by hydrogen bonds. An individual protein may contain both types of secondary structures. Some proteins, like collagen, contain neither but have their own more characteristic secondary structures. α Helix is a right-handed helix with 3.6 amino acid residues per turn. Hydrogen bonds are formed parallel to the helix axis. β Sheet is a parallel or antiparallel arrangement of the polypeptide chain. Hydrogen bonds are formed between the two (or more) polypeptide strands. β Turn or coils is a structure in which the polypeptide backbone folds back on itself. Turns are useful for connecting helices and sheets.

1.4 Protein Secondary Structure Prediction

Although secondary structure information alone is generally of only limited use, it is nonetheless helpful to be able to refer to a reliable secondary-structure prediction when attempting to predict the tertiary structure by fold recognition [7]. The relevance of secondary structure explains why secondary structure prediction from sequence has become one of the most ardently pursued tasks in bioinformatics. One idea is to use the particular secondary structure assignment is that (1) agrees most between proteins

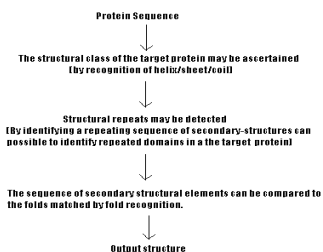
of similar structure, and/or (2) is the most predictable from sequence. Secondary structure prediction methods become increasingly important for prediction of general aspects of protein structure and function.

The amino acids typically found in α -helices differ considerably from those found in β -sheets. Alanine and Leucine often occur in α -helices, whereas Proline and Glycine are rare. In β -sheets Valine and Isoleucine are over-represented [8], whereas Glycine, Aspartic acid, and Proline are under-represented. Shorter structures such as 310-helices and β -bridges have distinct residue distributions. For 310-helices, the Alanine and Leucine signal has disappeared; instead the sequences are dominated by Proline, which often is observed as a helix initiator and breaker. For β -bridges, we no longer find a preference for Valine and Isoleucine. This finding indicates the role of the side chain in defining secondary and tertiary structure, an observation that can be built into new assignment methods provided based on Schulz, 1988; Fasman, 1989; Richardson and Richardson, 1989; Barton, 1995; Rost and Sander, 2000; Rost, 2001c.

1.5 Comparative Modeling

The protein-folding problem is one of the greatest remaining challenges in structural molecular biology (if not the whole of biology). Protein-structure prediction is, therefore, going to be vital to bridge the gap between structure and sequence determination. At present, the modeling of unknown protein structures by homology represents the best known method for protein-structure prediction.

Typically, structural biologists assume the protein fold to be the basic unit for structure classification. The fold and other basic structural elements are classified by automatic systems, such as SCOP, CATH, FSSP and MMDB [9]. When classified by experts, the particular features of a given fold are often described by the overall secondary structure arrangements, which therefore constitute a substantial step in protein classification. Functional aspects of proteins are also reflected in the secondary structure and occasionally function can be derived from secondary structure alone. There are four main uses of secondary structure: (1) it is indicative of the fold, (2) it influences the sequence alignment, (3) it is an intuitive means of visualizing protein structures, and (4) it is related to function.



1.6 Computational Biology

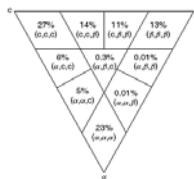
Since the arrival of information technology, bioinformatics has evolved from an interdisciplinary role to becoming a core program for a new generation of interdisciplinary courses [10]. Computational biology is the new interdisciplinary

subject that applies computer technology to solve biological problems and to manage and analyze biological information with programming. Every field of computational sciences including computational biology is evolving at such a rate that specific programs such as C, C++, PERL, Java applies on computational approaches to solve biological problems.

1.7 Colloc'h Triangle

Pattern recognition and machine learning approaches are considered to predict the structure of proteins, which plays an important role in predicting functional activities of a cell and tissues [11]. The Molecular Surface Package is a reimplementaion, in C for computing analytical molecular surfaces, areas, volumes, polyhedral molecular surfaces, and surface curvatures by Michael in 1993 [12]. The convex hull represented as a set of triangular faces is a close wrapping of the molecule. Each face belongs to one of the three classes of triangle patterns: small, large, or stretched allowing the depth of any molecular surface point to be defined [13, 14]. New strategies for the use of interactive computer graphics for man-machine communication in the field of molecular modeling s very helpful for the discussion of specific intermolecular interactions: attractive and repulsive forces towards an interaction partner can be mapped by color coding on the molecular surface.

Accurate assignments of secondary structures in proteins are crucial for a useful comparison with theoretical predictions. The overall number of residues in each of the three states (helix, strand or coil) differs on the number of helices or strands, thus implying a wide discrepancy in the length of assigned structural elements [15].



Colloc'h triangle

2 Chou-Fasman Algorithm

The Chou-Fasman algorithm for the prediction of protein secondary structure is one of the most widely used predictive schemes [16]. The Chou-Fasman method of secondary structure prediction depends on assigning a set of prediction values to a residue and then applying a simple algorithm to the conformational parameters and positional frequencies.

3 TertProt Algorithm

1. (Set Conformational Parameters) Set calculated propensities from a set of solved structures from Chou-fasman parameters/

2. (Calculation) The probabilities $pa[i]$, $pb[i]$, $pc[i]$ is determined as:

$$pa(i) = pa(j) \times pa(j+1) \times pa(j+2) \times pa(j+3)$$

$$pb(i) = pb(j) \times pb(j+1) \times pb(j+2) \times pb(j+3)$$

$$pc(i) = pc(j) \times pc(j+1) \times pc(j+2) \times pc(j+3)$$

3. (Prediction) If $pa(i)$ exceeds an arbitrary cutoff value equals or exceeds 1, alpha(or H) will be predicted. If $pb(i)$ exceeds an arbitrary cutoff value equals or exceeds 1, Beta(or B) will be predicted. If the first two conditions are not met the probability of a turn or coil (or C) will be predicted.

4. (Modelling) Using Protein secondary structure from the prediction, model a protein structure based on Colloc'h triangle.

4 Results

Four of the protein sequences were retrieved from the Protein Data Bank (PDB), a protein database and were analyzed for the secondary structure. The program has predicted the secondary structure containing Helices, Sheets and coils. Based on the prediction and using Colloc'h triangle, folds and three dimensional structures are predicted in TertProt method. The predictions have shown close relationship with the 3D structures available in the PDB database.

Fig. 1 to 4 also predicted that the edges are connected as the continuous blocks provided by Colloc'h triangle. Fig. 1 shows that BBB-BBC-BCC-CCC-CCH-CHH-HHH-HHH has been in a continuous graph path. The similar graph path has been also observed in Fig. 2, 3 and 4.

Peptide structures and sequences retrieved from Protein Data Bank are Peptide1:Oxytocin, Peptide2: HCMV protease, Peptide3: D,L altering peptide and Peptide4: Alzheimer beta amyloid

Protein 1:XY2-pdb Sequence: XY7QNCPLGX

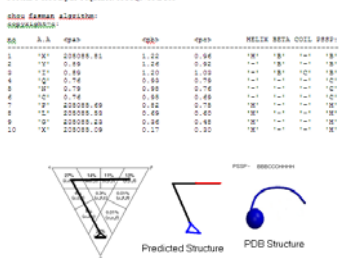


Fig. 1. Structure prediction of Oxytocin

Human cytomegalovirus (HCMV) protease. 1BFZ.pdb Sequence: XSYVKA

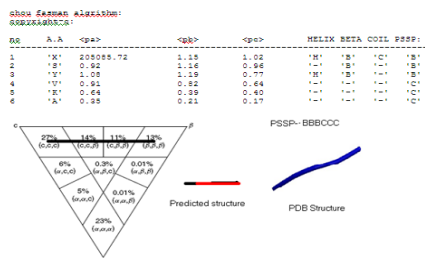


Fig. 2. Structure prediction for HCMV protease

Fig. 2 is the protein sequence retrieved from the Protein Data Bank with PDB ID: 1BFZ. The structure contains 6 amino acids and a peptide of Human cytomegalovirus protease. The sequence predicts with 3 sheets and 3 coils. Based on the arrangement of the secondary structure, a tertiary structure of the protein is constructed using Colloc'h triangle.

Fig. 3 is a D,L altering peptide with 9 aminoacids predicted with 6 peptides and 3 coils.

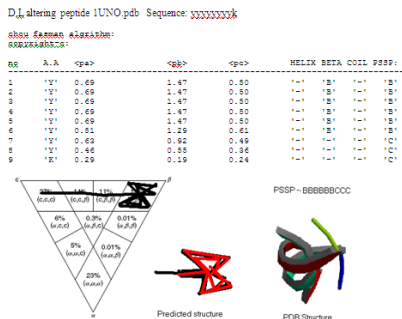


Fig. 3. Prediction of D,L altering peptide

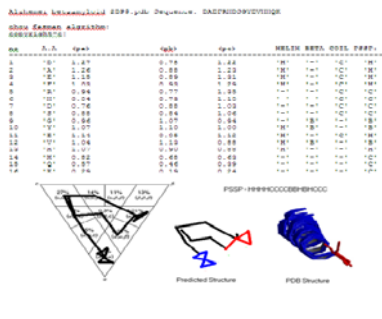


Fig. 4. Prediction in Alzheimer betaamyloid

Fig. 4 is a Beta amyloid peptide of Alzheimer. The structure contains folds due to bends in the path of Colloc'h triangle.

5 Discussion

The secondary structure is a key element of architectural organization of the secondary structure elements (SSE) (helix, strand, coil) in proteins and is an essential step for the analysis and modelling of protein structure [17]. However, Colloc'h et al. (1993) have shown some discrepancies between the assignments, pointing out difficulties in the assignment. This prompted them to propose an assignment computed from a ternary consensus method (TCM) used such as DSSP, DEFINE and P-CURVE based on distinct geometric criteria and algorithms.

Prediction of secondary structure of proteins provides information used to predict fold recognition and ab initio 3D structure of the proteins [18]. Colloc'h et al, 1993 probably the first researchers provided the concept of consensus with secondary structure assessment and prediction.

A new strategy for the use of interactive computer graphics for man-machine communication in the field of molecular modeling is very helpful for the discussion of specific intermolecular interactions: attractive and repulsive forces towards an interaction partner can be mapped by color coding on the molecular surface [19]. An ab initio method for building structural models of proteins from the scattering data has been already implemented by the computer program [20] A new tool has to be implemented through in silico methods for designing of three dimensional structures of proteins for better understanding of the architecture and functions of proteins. Comparative protein modeling is increasingly gaining interest since it is of great assistance during the rational design of a protein [21]. Ongoing genome sequencing and mapping projects have dramatically increased the number of protein sequences [22], which became importance to predict 3D structures of proteins.

6 Conclusion

Protein secondary structure prediction is one of the methods used to predict further ideas towards tertiary and quaternary structures. As NMR and X-ray crystallographic methods are costlier and time-taking methods, it is better to use computational approaches rather than conventional methods. The present approach provides the basic concepts of programming approach with logical predictions, which can provide the 3D protein prediction methods for proteins and peptides. Further programming approaches on TertProt have to be conducted in this area, which can provide information and regulation mechanisms in diseased proteins.

Acknowledgment. Author would like to thank management and staff of GITAM University, Visakhapatnam, India for their kind support in bringing out the above literature and providing lab facilities.

References

1. Wolfgang, K., Christian, S.: Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577–2637 (1983)
2. Dmitriy, F., Patrick, A.: Knowledge-Based Protein Secondary Structure Assignment. *Proteins: Structure, Function, and Genetics* 23, 566–579 (1995)
3. Floudas, C.A., Fung, H.K., McAllister, S.R., Mönnigmann, M., Rajgaria, R.: Advances in protein structure prediction and de novo protein design: A review. *Chemical Engineering Science* 61, 966–988 (2006)
4. Andrea, D.W., Leroy, H.: Systems Biology, Proteomics, and the Future of Health Care: Toward Predictive, Preventative, and Personalized Medicine. *Journal of Proteome Research* 3, 179–196 (2004)
5. Friedberg, E.C.: Biological Responses to DNA Damage: A Perspective in the New Millennium. *Quant. Biol.* 65, 593–602 (2000)
6. Howard, J.F., Christopher, W.V.H.: A fast method to sample real protein conformational space. *Proteins: Structure, Function, and Bioinformatics* 39, 112–131 (2000)
7. David, T.J.: A Practical Guide to Protein Structure Prediction. *Methods in Molecular Biology* 143, 131–154 (2000)
8. Tina, A.E., Linda, P., Janet, M.T.: Computational analysis of α -helical membrane protein structure: implications for the prediction of 3D structural models. *Protein Engineering, Design and Selection* 17, 613–624 (2004)
9. Fabrizio, C., Niccolò, T., Paul, M.W., Monica, B., Francesca, M., Massimo, S., Christopher, M.D.: Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nature Structural Biology* 6, 1005–1009 (1999)
10. Cooke, P.: The molecular biology revolution and the rise of bioscience megacentres in North America and Europe. *Environment and Planning C: Government and Policy* 22, 161–177 (2004)
11. Chen, Y., Blackwell, T.W., Chen, J., Gao, J., Lee, A.W., et al.: Integration of Genome and Chromatin Structure with Gene Expression Profiles To Predict c-MYC Recognition Site Binding and Function. *PLoS Comput. Biol.* 3, e63 (2007)

12. Michael, L.C.: The molecular surface package. *Journal of Molecular Graphics* 11, 139–141 (1993)
13. Anne, B.C., Julie, N., Laurent, B., Serge, H.: “Iso-depth contour map” of a molecular surface. *Journal of Molecular Graphics* 12, 162–168 (1994)
14. Franck, D., Jean-François, S., Jean-Paul, M.: Protein secondary structure assignment through Voronoi tessellation. *Proteins: Structure, Function, and Bioinformatics* 55, 519–528 (2004)
15. Colloc’h, N., Etchebest, C., Thoreau, E., Henrissat, B., Mornon, J.P.: Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. *Protein Eng.* 6, 377–382 (1993)
16. Rajbir, S., Sumandeep, K.D., Parvinder, S.S.: Chou-Fasman Method for Protein Structure Prediction using Cluster Analysis. *World Academy of Science, Engineering and Technology* 72, 982–987 (2010)
17. Labesse, G., Colloc’h, N., Pothier, J., Mornon, J.R.: P-SEA: a new efficient assignment of secondary structure from C trace of proteins. *CABIOS* 13, 291–295 (1997)
18. Joachim, S., Theo, M., Thomas, L.: Decicion tree-based formation of consensus protein secondary structure prediction. *Bioinformatics* 15, 1039–1046 (1999)
19. Brickman, J., Heiden, W., Vollhardt, H., Zachmann, C.D.: New man-machine communication strategies in molecular modeling. In: *Hawaii International Conference on System Sciences (HICSS 1995)*, p. 273 (1995)
20. Dmitri, I.S., Maxim, V.P., Michel, H.J.K.: Determination of Domain Structure of Proteins from X-Ray Solution Scattering. *Biophysical Journal* 80, 2946–2953 (2001)
21. Nicolas, G., Manuel, C.P.: SWISS-MODEL and the Swiss-Pdb Viewer: An environment for comparative protein modeling. *Electrophoresis* 18, 2714–2723 (1997)
22. Amos, B., Rolf, A.: The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucl. Acids Res.* 26, 38–42 (1998)

Analysis of *E.coli* Promoter Regions Using Classification, Association and Clustering Algorithms

D.S.V.G.K. Kaladhar¹, T. Uma Devi¹, P.V. Lakshmi², R. Harikrishna Reddy¹,
R.K. SriTeja Ayayangar V.¹, and P.V. Nageswara Rao²

¹ GIS, GITAM University, Visakhapatnam, India
{dkaladhar, nagesh.uma, harikrishnareddy.6, sriteja111}@gmail.com

² GIT, GITAM University, Visakhapatnam, India
pvlakshmi@gitam.edu, nagesh@gitam.in

Abstract. Data mining techniques can be well applied using various algorithms for the prediction of *E.coli* promoter regions. We studied various classification, Association and clustering algorithms like CART, Simple Logistic, BayesNet, Random forest, j48, LMT, Naïve Bayesian, Apriori and simpleKMeans over different *E.coli* promoter dataset. Random forest method using training dataset outperforms the remaining classification methods. The Association model (Apriori) predicted the presence of Adenine (A) at -45, -10 and -11 regions, Thiamine (T) at -35, -36 regions, Guanine (G) at -34 region. Cytosine (C) is not present in the submitted DNA data for *E.coli* promoter dataset at -14 to -9 and -36 to -31 regions using association model. Cluster based model using simpleKMeans predicted promoter regions true at -35 and -10 regions. If -36 to -31 region of the sequence contain TTGACA and -14 to -9 region contains TATAAT, there can be highest probability of finding promoter in *E.coli*. The condition becomes false, if the -36 to -31 region contains ACGACG and -14 to -9 contain TGAATG.

Keywords: Data mining, Algorithms, *E.coli* promoter.

1 Introduction

Research on Probability and statistics mainly focuses on distributions and hypothesis testing. Data can be used to estimate the parameters of the model, and then delivers the prediction and classification of the data for evidential support. Scientific data provides a platform to learn from the data and make true predictions [1]. In general, data mining is the search for hidden patterns that may exist in large databases [2]. The primary task in data mining is the development of models about aggregated data [3].

The combined efforts of biologists, statisticians, mathematicians, computer scientists, and software engineers made an understanding of both the biology and the computational methods, which are essential for tackling the associated ‘data mining’ tasks [4]. In contrast to pattern discovery and data mining, class prediction methods are techniques explicitly designed to classify objects into known groups. Computational techniques for classifying multidimensional data are well described in machine-learning literature.

Bioinformatics provides opportunities for mounting novel data mining methods, is becoming a common complement to many scientific areas like medicine and biotechnology [5], [6]. The scientists in 21st century is Exploring and analyzing the vast volumes of data compared to the past. Information visualization and visual data mining can help to deal with the flood of information and to discover heretofore unknown information [7], [8].

An effective method in stepwise representation of instructions for calculating a function is called an algorithm. Various statistical tests will be applied to determine whether one learning algorithm outperforms another on a particular learning task [9]. Classification has been studied extensively in the past with limiting their suitability for data mining large data sets [10].

The tasks and applications of data mining are broad and diverse with flourishing research and development activities and successful systems report [11]. Data mining may be strongly influenced in the design of data mining languages. Data mining will be an emerging field and is expected that various kinds of flexible, interactive user interfaces can be done by the computational scientists [12].

Data mining aims at the construction of semi-automatic tools for the analysis of large sets of information [13]. Data Mining is a step in the KDD (Knowledge Discovery in Databases) process consisting of classification, data analysis and discovery algorithms that, under adequate computational efficiency limitations, produce a particular record of patterns over the data.

There is a wide variety of data mining algorithms and some of the most popular and references therein can be seen in Fayyad et al., 1996 [14], [15]. Popular data mining algorithms include Nearest Neighbor technique, Naïve Bayesian network, Tree-Augmented Network Classification technique and a decision tree classification technique based on the ID3-algorithm etc

Researchers in many different fields have shown great interest in data mining as a key research topic in database systems and machine learning and by many industrial companies as an important area with an opportunity of major revenues [16]. Classification is a data mining (machine learning) technique used to predict group membership for data instances and includes analysis and prediction [17].

The vital learning methods in Waikato Environment for Knowledge Analysis (WEKA) are “classifiers”, and they induce a rule set or decision tree that models the data [18], [19].

DNA contains four kinds of nucleotides adenine (A), guanine (G), cytosine (C), and thymine (T) that encodes genetic information. The DNA sequence of 168 promoter regions (-150 to +10) for *Escherichia coli* (or *E.coli*) RNA polymerase had been defined by genetic (promoter mutations) or biochemical (5' end determination) criteria have been shown to contain two regions of conserved DNA sequence, located about 10 and 35 base pairs upstream of the transcription start site [20].

Borries and Guangwen, 1991 were trained on a set of 80 known promoter sequences combined with different numbers of random sequences using neural network. The conserved -10 region and -35 region of the promoter sequences and a combination of these regions were used in three independent training sets predicted 100% accuracy on the promoter test set [21].

In this paper, we studied various classification, Association and clustering algorithms like CART, Simple Logistic, BayesNet, Random forest, j48, LMT, Naïve

Bayesian, Apriori and SimpleKMeans over different *e.col* promoter dataset. Classification is the main objective to estimate the performance of these algorithms over *E. coli* promoter dataset.

2 Classification Models Tested

2.1 CART

The algorithm is based on Classification and Regression Trees (CART) by Breiman et al (1984). A CART tree is a binary decision tree that produces binary splits at each node [22].

2.2 Simple Logistic

Logistic models use linear logistic regression with simple regression functions as base learners is used for fitting the logistic model for the data [23, 24]. The WEKA implementation is called SimpleLogistic.

2.3 BayesNet

Bayesian classifiers try to model probabilistic relationships between the attributes and the class variable. It uses the well-known Bayes theorem to combine a priori information with evidence from data. Let X denote the attribute set and Y denote the class variable. The classifier learns the posterior probability $P(Y | X)$ for every combination of X and Y so that a new record can be classified such that the posterior probability is maximal. This technique is used in different ways for different algorithms, where BayesNet is an implementation of a BayesNet.

2.4 Random Forest

Random forest is an ensemble method, which means that it aggregates the prediction of multiple classifiers to improve accuracy. The random forest algorithm selects a subset of the input features as training set, and uses decision trees as its base classifier. The resulting decision trees are combined to get a good accuracy.

2.5 j48

j48 is a standard classification algorithm that is widely used for practical machine learning. This is an implementation of C4.5 release 8 which produces decision trees.

2.6 LMT

A Logistic Model Tree (LMT) is an algorithm used for supervised learning tasks that is combined with linear logistic regression and tree induction methods. LMT creates a model tree with a standard decision tree structure with logistic regression functions at leaf nodes and has an associated logic regression functions instead of just class labels.

2.7 Naïve Bayesian

Naïve Bayesian classifier is a simple probabilistic classifier based upon Bayes theorem with strong (naive) independence assumptions. The advantage of this algorithm is that it rebuild amount of training data to estimate the parameters (means and variances of the variables) necessary for classification.

3 Association and Clustering Models Tested

3.1 Apriori

Mining of association rules has received a lot of attention in recent years by various computational scientists. Comparing with other machine learning techniques, the main advantage is a low number of database passes done when searching the hypothesis space One of the best-known association rule learning algorithms is Apriori [25], [26]. This algorithm was extensively studied and successfully applied in many problem domains, adapted to other areas of machine learning and data mining techniques[27], [28].

3.2 SimpleKMeans

SimpleKMeans clustering [29] is one of the simplest unsupervised learning algorithms that solve the well known clustering problems such as prediction of promoter regions in *E.coli* . The main idea is to define k centroids placed in a cunning way which causes different result.

4 Research Methodology

Several algorithms can be examined for goodness of fit, based on the datasets submitted to available tools. The purpose of sample structure is to demonstrate the prediction of promoter regions in the *E.coli* The data set contains 106 instances and 59 attributes. The first attribute contains +/- indicating the class promoter. The second attribute contains the name of the promoter. The remaining 57 fields are the sequence, starting at position -50 (p-50) and ending at position +7 (p7). Each of these fields is filled by one of {a, g, t, c}. The algorithms such as CART, Simple Logistic, BayesNet, Random forest, j48, LMT and Naïve Bayesian, were evaluated for their performance and the results from this analysis are discussed.

5 Experimental Results

The analytical algorithms used with *E.coli* promoter data are from the data mining software suite WEKA. They are particularly suitable when examining continuous time dependent attribute value vectors. In particular the implementations of classification algorithms for CART, Simple Logistic, BayesNet, Random forest, j48, LMT and Naïve Bayesian were tested.

Table 1 provides the output for comparison of the classification techniques including CART, Simple Logistic, BayesNet, Random forest, j48, LMT and Naïve

Bayesian over different *E.coli* promoter dataset. Machine learning approach shows that Random forest method using training dataset outperforms the remaining methods.

Tree generated by CART Classifier for the classification of dataset is shown in Figure 1. There is one leaf node predicted using CART. The size of the tree predicted by CART contains one branch.

Table 1. Classification models for *E.coli* promoter using training dataset

S.No	Algorithm	Data using Training dataset		
		Correctly Classified	Incorrectly Classified	Absolute relative error
1	CART	32	68	99.9%
2	Simple Logistic	46	54	92
3	BayesNet	91.5	8.5	13.3
4	Random Forest	100	0	24.8
5	j48	83	17	31.8
6	LMT	46	54	92
7	Naive Bayesian	91.5	8.5	16

```

CART Decision Tree
: t(34.0/72.0)

Number of Leaf Nodes: 1

Size of the Tree: 1
    
```

Fig. 1. Classification model using CART Decision Tree

Figure 2 provides the best rules predicted based on the submitted *E.coli* promoter datasets. The model predicted the presence of Adenine (A) at -45, -10 and -11 regions, Thiamine (T) at -35, -36 regions, Guanine(G) at -34 region and prediction of Cytosine (C) are not present in the submitted DNA data for *E.coli* promoter dataset.

```

Best rules found:

1. p-45=a p-35=t 30 ==> class+= 30 conf:(1)
2. p-35=t p-34=g 35 ==> class+= 34 conf:(0.97)
3. p-45=a p-36=t 28 ==> class+= 27 conf:(0.96)
4. p-35=t p-10=a 27 ==> class+= 26 conf:(0.96)
5. p-36=t p-35=t p-34=g 27 ==> class+= 26 conf:(0.96)
6. p-36=t p-34=g 34 ==> class+= 32 conf:(0.94)
7. class+= p-45=a 32 ==> p-35=t 30 conf:(0.94)
8. p-36=t p-10=a 30 ==> class+= 28 conf:(0.93)
9. class+= p-10=a 30 ==> p-36=t 28 conf:(0.93)
10. p-35=t p-11=a 29 ==> class+= 27 conf:(0.93)
    
```

Fig. 2. Association data model using Apriori

Figure 3 provided Cluster based model using simpleKMeans. The result presented 0/(false) and 1/(True) data predicted in the cluster of dataset. The promoter regions predicted true at -35 and -10 regions upstream of the transcription start site (TATTA/TTTTT at 1 to 5). If -36 to -31 region of the sequence contain DNA string “TTGACA” and -14 to -9 region contain DNA string “TTTAAT”, there can be highest probability of finding promoter in *E.coli*. The condition is also true (highest probability of finding promoter in *E.coli*), if -36 to -31 region of the sequence contain DNA string “TTGACA” and -14 to -9 region contain DNA string “TATAAT”. The condition again may be false, if the -36 to -31 region of promoter in *E.coli* contains DNA string “ACGACG” and -14 to -9 regions (from the start) contains DNA string “TGAATG”. Hence the regions of promoter in *E.coli* are well predicted based on True (1)/False (0) results placed on cluster analysis.

Cluster centroids:

Attribute	Cluster#			p-25	g	a	g
	Full Data (106)	0 (47)	1 (59)				
				p-24	g	a	g
				p-23	t	t	c
				p-22	t	g	t
class	+	-	+	p-21	a	a	a
instance	1019	1019	1024	p-20	a	t	a
p-50	t	t	t	p-19	a	g	a
p-49	a	g	a	p-18	t	c	t
p-48	a	a	g	p-17	t	t	g
p-47	c	c	g	p-16	a	g	c
p-46	a	g	a	p-15	t	t	t
p-45	a	a	a	p-14	t	t	t
p-44	a	a	a	p-13	t	g	a
p-43	a	a	t	p-12	t	a	t
p-42	a	t	a	p-11	a	a	a
p-41	a	c	a	p-10	a	t	a
p-40	a	a	a	p-9	t	g	t
p-39	c	a	c	p-8	t	a	t
p-38	t	a	t	p-7	t	t	c
p-37	c	t	c	p-6	g	t	g
p-36	t	a	t	p-5	c	a	c
p-35	t	c	t	p-4	c	t	c
p-34	g	g	g	p-3	c	t	c
p-33	a	a	a	p-2	c	a	c
p-32	c	c	c	p-1	c	a	c
p-31	a	g	a	p1	c	t	c
p-30	t	a	t	p2	t	c	a
p-29	a	g	a	p3	t	a	t
p-28	g	c	g	p4	c	t	c
p-27	t	t	g	p5	c	t	a
p-26	t	g	a				

Fig. 3. Cluster based model using SimpleKMeans

6 Discussions

Researchers from Bioinformatics and computational sciences are frequently facing problem of evaluating the accuracy of the particular prediction algorithm. Some of the methods are optimized to produce very few false positives. The prime interest of prediction algorithm is to perform novel data that have not been used in the process of constructing an algorithm from the same data domain.

Evaluation and redundancy of the data will be used in problem solving of the training and testing a particular algorithm for predicting accuracy of the data. There are several performance evaluation algorithms to evaluate degree of similarity between training and test data in prokaryotic and eukaryotic organisms [30]. Cluster analysis algorithms are mostly used for data reduction and classification of objects.

Information metabolism such as the development of cellular structure, communication and regulation, provides a way to store and retrieve the information that guides the metabolic process of life. Like other metabolic pathways, this process of life is highly regulated in *E.coli*, a prokaryotic bacteria attached to human life. Information can be transferred from one generation to another in the form of DNA and the processes are limited to specific portions of the cell cycle. According to central dogma of life, information of DNA replication and meiosis is retrieved by the transcription of DNA into RNA and the ultimate translation of the signals in the mRNA into protein [31].

The nucleotide sequence can be directly preceding the start site, is strongly homologous to the prokaryotic promoter consensus sequence. Analysis of DNA sequences from -50 to +10, has shown known transcriptional start points for genes of *E. coli*. Alignment of the complete list used for reference until successive analyses did not alter in the structure of the list. The final compilation of all bases in the -35 (TTGACA) and -10 (TATAAT) hexamers were highly conserved. 92% of promoters had inter-region spacing of 17 ± 1 bp, and 75% of the uniquely defined start points initiated 7 ± 1 bases downstream of the -10 region [32].

7 Conclusion

RNA polymerase recognizes its promoters through base-specific interaction between defined segments of DNA and the σ subunit of the enzyme in *E.coli*. A novel computer procedure using machine learning has been used to search for prediction among *E.coli* promoter sequences.

Acknowledgments. Authors would like to thank management and staff of GITAM University, Visakhapatnam, India for their kind support in bringing out the above literature and experimentation.

References

1. Hastie, T., Tibshirani, R., Friedman, J., Franklin, J.: The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer* 27, 83–85 (2005)
2. Ng, R.T., Han, J.: Efficient and Effective Clustering Methods for Spatial Data Mining. In: Proc. 20th VLDB Conf., Santiago, Chile, pp. 144–155 (1994)
3. Agrawal, R., Srikant, R.: Privacy-preserving data mining. *ACM SIGMOD Record* 29, 439–450 (2000)
4. Slonim, D.K.: From patterns to pathways: gene expression data analysis comes of age. *Nature Genetics Supplement* 32, 502–508 (2002)
5. Houston, A.L., Chen, H., Hubbard, S.M., Schatz, B.R., Ng, T.D., Sewell, R.R., Tolle, K.M.: Medical Data Mining on the Internet: Research on a Cancer Information System. *Artificial Intelligence Review* 13, 437–466 (1999)

6. Cios, K.J., Moore, G.W.: Uniqueness of medical data mining. *Artificial Intelligence in Medicine* 26, 1–24 (2002)
7. Keim, D.A.: Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics* 8, 1–8 (2002)
8. Hearst, M.A.: Untangling text data mining. In: Proc. 37th Ann. Meeting Assoc. for Computational Linguistics (ACL 1999), pp. 3–10 (1999)
9. Dietterich, T.G.: Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation* 10, 1895–1924 (1998)
10. Mehta, M., Agrawal, R., Rissanen, J.: SLIQ: A Fast Scalable Classifier for Data Mining. In: Apers, P.M.G., Bouzeghoub, M., Gardarin, G. (eds.) *EDBT 1996*. LNCS, vol. 1057, pp. 18–32. Springer, Heidelberg (1996)
11. Han, J., Fu, Y., Wang, W., Koperski, K., Zaiane, O.: DMQL: A data mining query language for relational databases. In: Proc. of the SIG-MOD 1996 DKMD Workshop, Montreal, Canada, pp. 27–33 (1996)
12. Piatetsky-Shapiro, G., Frawley, S.J.: *Knowledge Discovery in Databases*. AAAI Press/MIT Press, Menlo Park, CA (1991)
13. Cheng, C.H., Fu, A.W., Zhang, Y.: Entropy-based subspace clustering for mining numerical data. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-1999), pp. 84–93 (1999)
14. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: Knowledge Discovery and Data Mining: Towards a Unifying Framework. In: Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD 1996), Portland, Oregon, pp. 82–88 (1996)
15. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery: An overview. In: Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.) *Advances in Knowledge Discovery and Data Mining*, pp. 1–30. AAAI Press, Menlo Park (1996)
16. Chen, M.S., Han, J., Yu, P.S.: Data mining: an overview from a database perspective. *Knowledge and Data Engineering* 8, 866–883 (1996)
17. Witten, I.H., Frank, E.: *Data Mining: Practical machine learning tools and technique*, 2nd edn. Morgan Kaufmann, San Francisco (2005)
18. Witten, I.H., Frank, E.: *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco (2000)
19. Agrawal, R., Imielinski, T., Swami, A.N.: Database mining: a performance perspective. *IEEE Trans. Knowledge and Data Engineering* 5, 914–925 (1993)
20. Hawley, D.K., McClure, W.R.: Compilation and analysis of *Escherichia coli* promoter DNA sequences. *Nucl. Acids Res.* 11, 2237–2255 (1983)
21. Demeler, B., Zhou, G.: Neural network optimization for *E.coli* promoter prediction. *Nucl. Acids Res.* 19, 1593–1599 (1991)
22. Breiman, L., Friedman, J., Olshen, R.A., Stone, C.J.: *Classification and Regression Trees*. Wadsworth International Group, Belmont (1984)
23. Landwehr, N., Hall, M., Frank, E.: Logistic Model Trees. *Machine Learning* 59, 161–205 (2005)
24. Sumner, M., Frank, E., Hall, M.: Speeding up Logistic Model Tree Induction. In: Jorge, A.M., Torgo, L., Brazdil, P.B., Camacho, R., Gama, J. (eds.) *PKDD 2005*. LNCS (LNAI), vol. 3721, pp. 675–683. Springer, Heidelberg (2005)
25. Agrawal, R., Imielinski, T., Srikant, R.: Mining association rules between sets of items in large Databases. In: *Proceedings of ACM SIGMOD Conference on Management of Data*, pp. 207–216 (1993)

26. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: Proceedings of the 20th International Conference on Very Large Databases, Santiago, Chile, pp. 487–499 (1994)
27. Bayardo, R.J., Agrawal, R., Gunopulos, D.: Constraint-based rule mining in large, dense databases. In: Proceedings of the 15th International Conference on Data Engineering, pp. 188–197 (1999)
28. Megiddo, N., Srikant, R.: Discovering predictive association rules. In: Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (KDD 1998), pp. 274–278 (1998)
29. Lin, D., Wu, X.: Phrase Clustering for Discriminative Learning. In: Proceedings of ACL-IJCNLP 2009 (2009)
30. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A.F., Nielsen, H.: Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* 16, 412–424 (2000)
31. Gilbert, H.F.: *Basic concepts in Biochemistry: A student's survival guide*, 2nd edn., vol. 48. McGraw-Hill, Health Professions Division, USA (2000)
32. Harley, C.B., Reynolds, R.P.: Analysis of *E.Coli* Promoter sequences. *Nucl. Acids Res.* 15, 2343–2361 (1987)

A Parameterization Study of Short Read Assembly Using the Velvet Assembler

Alex Christopher Elliot, A. Louise Perkins, and Sumanth Yenduri

University of Southern Mississippi,
730 East Beach Blvd, Long Beach, MS 39560

Abstract. In this study, we examine approaches to the problem of assembling large, contiguous sections of genetic code from short reads generated from laboratory techniques. We explore the Eulerian Path approach in detail, utilizing a de Bruijn Graph, and demonstrate current software technologies and algorithms using a sample genome. We investigate the input parameters of Velvet and discuss their implications.

Keywords: Velvet, De Bruijn Graphs, Genetic Assembly, Euler Path.

1 Introduction

Since the discovery of the DNA double helix in 1953 [1], science has sought to fully understand the information contained within it [2]. In a macro view, understanding an organism's genome can help reveal its phylogeny and origins, while the micro view can uncover information about disease susceptibility and cure. Small sections of an individual organism's genetic fingerprint that indicate the presence or absence of a particular trait are called genetic biomarkers. These biomarkers can be used to, for example, determine relation between organisms, gauge exposure to a particular genetic toxicant, predict inherited disease, or determine an optimal treatment approach. In order to understand genetic information, one must find a way to read the information contained within DNA or RNA. Genetic sequencing techniques were first developed in the early 1970's [3]. These complex methods, including the wandering-spot technique were very labor intensive.

Fredrick Sanger [4] and Gilbert [3] independently published research in 1977 that greatly simplified the sequencing process. The Sanger method is a chain terminating technique that uses of dideoxynucleotide triphosphates (ddNTPs) to selectively terminate long strands of genetic material [5]. In this method, single stranded, denatured DNA source material is cloned and separated into four separate solutions containing one of ddATP, ddTTP, ddCTP, or ddGTP each. The dideoxynucleotides terminate the multiple copies of the DNA strand at each location of the target base, resulting in strands that begin at the origin and have length of the base location index. The output of the four dideoxynucleotide solutions is then separated by gel electrophoresis or fluorescent absorption if dyes were used. The result is an index location of each base in the source DNA to a one-base resolution. Sanger sequencing

generates long reads of about one thousand bases, but requires weeks to months of costly laboratory time [6]. This technique is susceptible to cloning error [7], as parts of the cloning vector may enter the resulting sequence.

An alternative to Sanger sequencing, pyrosequencing was developed by Nyrén and Ronaghi at the Royal Institute of Technology in Stockholm in 1998 [8]. This method involves iterative addition of bases in an enzymatic solution of Sulfurylase, Luciferase and Apyrase. As each base bonds to the source material, a measurable amount of light is released per base. Repeat bases yield proportionately more light. After each base is introduced and bound, an enzyme is added to remove all unused bases before the next base is added.

Pyrosequencing results in short length reads with an upper limit of approximately 500 bases, however commercial implementations are constantly increasing the maximum read length. Pyrosequencing is also less expensive to perform than traditional techniques, with companies such as 454 Life Sciences producing all-in-one units [9]. This technique can produce approximately 25Mbp/4hr [10]. As this technique does not require traditional cloning, it is not susceptible to vector cloning error. It is, however, potentially less accurate in homopolar regions with a long series of repeating bases. Pyrosequencing techniques normally result in many copies of overlapping short reads. After the laboratory work is complete, the reads must then be assembled into a representation of the source sequence. Although various solutions exist for this problem, all require some amount of a priori assumption and reliance on yet to be fully verified metrics. The challenge, algorithmically, is to determine how each of the reads fits into the larger sequence. Information to support the selection amongst candidate solutions can come from existing, reference genomes, statistical models, or sheer read coverage. Once sequenced, data can be added to large, publically accessible genome databases such as NCBI [11] or GenomeNet [12]. The NCBI Basic Local Alignment Search Tool (BLAST) can be used to find regions of local similarity between sequences. The program [13] compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

2 The EULERIAN Approach

In this section, we describe the application of the EULERIAN path to short read assembly and its differences as compared to earlier methods. We discuss one available implementation – Velvet, and provide insight into its algorithm. Older approaches to the problem of read assembly were designed around the assembly of few, long reads. Many available programs utilized the “overlap-layout-consensus” paradigm which tests each possible read pair combination to determine the best matches. Each read is represented as a node, and each detected overlap is drawn as an arc between the overlapping nodes. Once matches are scored, the assembly is generated based on overlap scoring. Unfortunately, determining the layout leads to the NP-complete Hamiltonian Path Problem [14]. The difficulty of the Hamiltonian Path Problem is exacerbated when attempting to operate on an increased number of reads.

Pevzner proposed an alternative solution to the read assembly problem for sequencing by hybridization [15]. By making use of the de Bruijn Graph, he reduced read assembly to a solvable Eulerian Path Problem. Further work by Idury and Waterman [16] applied the Eulerian path to short fragment assembly by treating short fragment assembly as Sequencing by Hybridization problem. Pevzner, Tang and Waterman refined their Eulerian graph techniques in 2001 to include methods of error correction and repeat handling in data [15]. A de -Bruin graph is a directed, n -dimensional graph of m symbols that represents overlaps between sequences of symbols. In graph theory, an n -dimensional de Bruijn graph of m symbols is a directed graph representing overlaps between sequences of symbols. It has mn vertices, consisting of all possible length- n sequences of the given symbols (the same symbol may appear multiple times in a sequence). If one of the vertices can be expressed by shifting all symbols by one place to the left and adding a new symbol at the end of another vertex, then the latter has a directed edge to the former vertex. Although de Bruijn graphs are named after Nicolaas Govert de Bruijn, they were discovered independently both by de Bruijn (1946) and I. J. Good (1946). Much earlier, Flye Sainte-Marie (1894) implicitly used their properties.

Zerbino and Birney released a set of algorithms called “Velvet” [17] to manipulate de Bruijn graphs for genomic sequence assembly. In their implementation of the graph, a k -mer is defined as a substring of length k , extracted from a read. Each node contains a series of overlapping k -mers, with each overlap having length $k-1$ bases. Each node is attached to another, “mirror” node which contains the reverse series of k -mers. These mirror nodes take into account the complementary nature of genetic material. Nodes whose last k -mer overlaps with the first k -mer of another node are connected by a directed arc. (Fig 2.1) The assembled contiguous sequence or “contig” is represented by a traversal from the first k -mer of the first node through connected arcs to each other node.

Once the input reads have been hashed into k -mers and assembled into nodes and arcs, the resulting graph must be simplified and cleared of errors. Velvet simplifies the graph by combining adjacent nodes with only one incoming and outgoing arc. This reduces the node count to only nodes with multiple arcs. Error correction is performed by eliminating “tips” and “bulges.” A “tip” is defined as a chain of nodes connected at only one end, and Velvet removes tips that do not meet minimum length and coverage requirements. A “bulge” is a redundant path that starts and ends at the same nodes as other paths with similar sequences. Velvet again employs a length threshold and simple sequence identity to condense or merge a bubble. Velvet is thus composed of four stages: hashing the reads into k -mers, constructing the de Bruijn graph, correcting errors, and resolving repeats. The first stage, graph construction, is memory intensive. The time complexity of error correction depends mainly on number of nodes in the graph, which is a result of read coverage, error rate, and number of repeats in the source material. The graph search used during error detection and correction employs the Dijkstra algorithm which has a time complexity of $O(N \log N)$ when implemented with a Fibonacci heap [18]. Repeat resolution also depends on the number of nodes present in the graph and the average length of those nodes.

3 Methods

To illustrate the operation of Velvet, we chose a specific, active coding gene of *Escherichia coli* str. K-12 substr. MG1655. This gene, NP_415534, codes for the enzyme proline dehydrogenase/pyrroline-5-carboxylate dehydrogenase which functions as a fused DNA-binding transcriptional regulator [12]. The *E. coli* genome has been extensively studied and fully sequenced [19] allowing for comparison of our assembly results with established sequence data. The NP_415534 gene sequence was obtained from GenomeNet [12] in its full form as an ASCII formatted fasta file [11]. This reference gene contains a total of 3963 ordered nucleotides.

From the reference file, we used the read simulation function of MetaSim [20] to output two sets of simulated reads. The first set represents an “exact” or reference set in which, 5000 reads were taken directly from the source gene without introduced error. The output reads have a normal distribution across the source gene and an average read length of 997.87 base pairs. To illustrate real world data, we also generated a set of reads modeling the read output of the LifeSciences 454 sequencer [9]. These 5000 simulated reads contained 29890 insertions and 7321 deletions. Each insertion is the addition of an extra base not present in the original material. Each deletion is the removal of one base from the original material. Locations of these induced errors are based on characteristics of pyrosequencing such difficulties accurately reading homopolar regions. Average Read Length was 258.21 base pairs. Each of the simulated read sets were run through the Velvet Assembler using varying values of k-mer length (k), expected coverage (exp_cov) and coverage cutoff (cv_cut). Automation of parameter variation and report generation was assisted by the standardized velvet assembly report script project [21]. Expected coverage is the expected frequency of repeats of each source base. This is a function of the source material and the depth at which the sequencing was performed.

Table 1 shows the parameter permutations used and their results for the simulated 454 reads. “kmer” is the selection of k or kmer length. “cvCut” is coverage cutoff, a threshold used to determine if a node in the constructed de Bruijn graph should be included as part of the final assembly. “exp,” expected coverage, is the expected frequency of repeats of each source base. “ctgs” is the number of contigs. “asmLg,” “mean,” and “max” refer to the total length, mean, and maximum length of all assembled contigs respectively. “N50” refers to the length of the shortest contig in an assembly such that the sum of contigs of equal length or longer is at least 50% of the total length of all contigs. “1k” is the number of contigs over 1000 bases long. “tiles” is the number of reads that are used in an assembly. “rdPc” is percentage of input reads used in the assembly. Lower frequency nodes with coverage below the coverage cutoff value are suspected to be erroneous and are subsequently removed during graph error correction, especially during tip and bulge removal. This threshold specifies how many read k-mers must overlap for each contig kmer. The number of kmers per read is a function of read length L and k-mer length K (L-K+1) [21]. AMOS files of selected final assemblies were generated with velvet and opened for analysis with Hawkeye [22].

Table 1. Assembly Parameter Permutations

kmer	cvCut	exp	ctgs	asmLg	N50	mean	lk	max	files	rdPc
21	2	4	42593	114471	26	0	56	1394	27.88	
21	3	6	2603	70464	26	0	44	2236	44.72	
21	4	8	1484	40732	27	0	70	2638	52.76	
21	6	12	107	3896	36	0	144	3375	67.50	
21	10	20	60	4007	91	66	0	381	4997	99.94
21	12	24	56	4366	110	77	0	411	5000	100.00
23	2	4	4337	128149	29	29	0	60	1378	27.56
23	3	6	2638	78246	28	29	0	66	2914	58.28
23	4	8	692	20845	29	30	0	57	1438	28.76
23	6	12	212	7328	35	34	0	119	3065	61.30
23	10	20	65	4453	99	68	0	311	4999	99.98
23	12	24	46	4099	133	89	0	344	4998	99.96
27	2	4	3989	137723	33	34	0	82	1655	33.10
27	3	6	2418	83130	32	34	0	62	1588	31.76
27	4	8	1067	37296	34	34	0	81	2490	49.80
27	6	12	253	9410	41	40	0	130	3637	72.74
27	10	20	47	4250	136	90	0	559	4999	99.98
27	12	24	44	4331	145	98	0	594	5000	100.00
31	2	4	3804	150919	38	39	0	81	1118	22.36
31	3	6	1251	49497	37	39	0	73	596	11.92
31	4	8	383	15473	39	40	0	88	1053	21.06
31	6	12	134	7062	54	52	0	195	4228	84.56
31	10	20	22	3605	266	163	0	497	4975	99.50
31	12	24	22	3790	273	172	0	491	5000	100.00

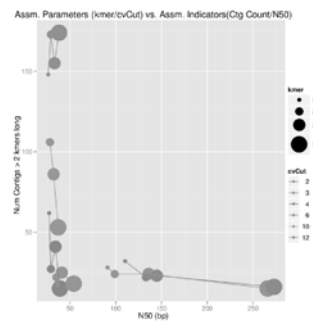
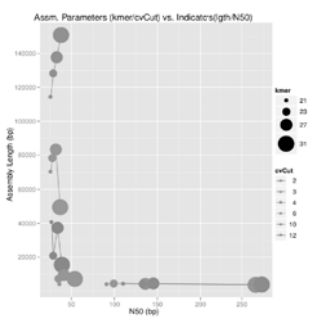


Fig. 1. Assembly Parameters (kmer/cvCut) vs. Indicators (lgth/N50). This scatterplot illustrates the effect of kmer length and coverage cutoff on N50 and assembly length. N50 refers to the length of the shortest contig in an assembly such that the sum of contigs of equal length or longer is at least 50% of the total length of all contigs. Here we see a logarithmic distribution where higher cvCut values generate larger contigs.

Fig. 2. Assembly Parameters (kmer/cvCut) vs. Assembly Indicators (Ctg Count/N50). This plot shows the influence of kmer length and cvCut on the number of contigs produced with greater than 2*k length. Again, we see a somewhat logarithmic function with higher cvCut and higher kmers producing longer and fewer isolated contigs.

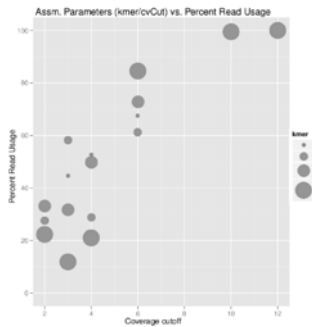


Fig. 3. Assembly Parameters (kmer/cvCut) vs. Percent Read Usage. This figure compares kmer and cvCut to the percent read usage. With sufficiently high k, read utilization increases with coverage cutoff.

The results of these assemblies were then compared back to the original, reference gene sequence using BLASTN [13]. BLASTN outputs a percent identity which shows the similarity to the reference sequence as well as a base by base alignment which shows direct matches, deletions, insertions and substitutions for each assembled node.

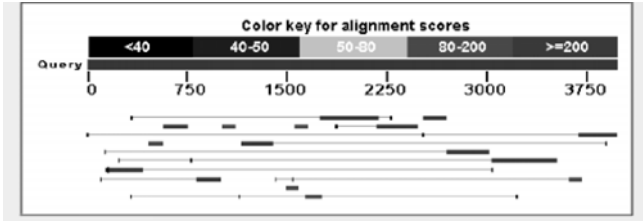


Fig. 4. BLAST Map for Simulated 454 Reads ($k = 31$, coverage cutoff = 12, expected coverage = 24). BLAST maps the 16 resultant contigs of the 454 simulated reads at k -mer length 31, expected coverage 24, and coverage cutoff 12. The level of similarity to the reference gene is shown by the color, with red being the best quality. Contigs appear to high quality and well distributed.

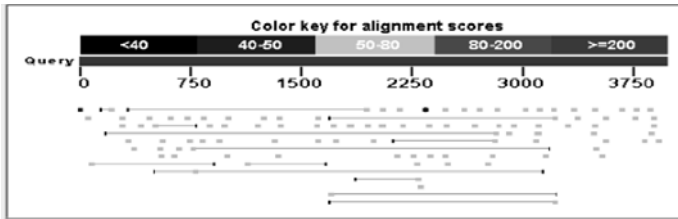


Fig. 5. BLAST Contig Scoring for Simulated 454 Reads ($k = 21$, coverage cutoff = 2, expected coverage = 4). This assembly resulted in 4253 nodes and $n50$ of 16. BLAST maps the resultant contigs above a scoring threshold. The level of similarity to the reference gene is shown by the color, with red being the best quality. Contigs appear to be very small with medium to poor quality, yet well distributed.

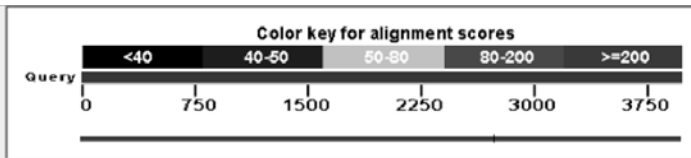


Fig. 6. BLAST Contig Scoring for Simulated “Exact” Reads ($k = 21$, coverage cutoff = 2, expected coverage = 4). This assembly maps the single resultant contig of the simulated “exact” reads at k -mer length 21, expected coverage 4, and coverage cutoff 2. The assembled contig achieved 100% reference gene coverage at 100% identity with a length of 3963 bases.

4 Conclusions

Read assembly remains an inexact science, relying heavily on statistical modeling and inference for error correction and graph simplification. Our synthetic assembly

experiment demonstrates how heavily parameter selection influences final assembly, thus consideration must be made when designing an experiment and performing the assembly. The value of k depends primarily on the nature of the source genome, particularly the length and abundance of repeats. With sufficiently high k , read utilization and resultant contig length increases with coverage cutoff, due to the removal of lower coverage nodes, however this elimination can lead to mis-assemblies. A delicate balance exists between easing coverage limits to increase final assembled contig length and a reduction in accuracy. Some experiments, such as preliminary genome sequencing may seek wider coverage and fewer but longer nodes at the expense of 100% accuracy of individual bases, whereas small target sequencing of short gene segments may obtain the higher accuracy required by increasing read coverage. As the algorithms continue to mature, research into the automated choice of parameters will assist scientists when faced with the challenge of read assembly. Obtaining and integrating the various scripts and applications was a chore, as each had its own set of dependencies and special setup instructions. Velvet assembly and the associated tools would benefit from a cloud implementation, similar to that of NCBI's BLAST to provide a full suite of assembly tools with minimal or no configuration. Further efforts to understand the parameterization of short read assembly using Velvet should expand both the source data and selected parameter value set, possibly to include eukaryotic data. A more detailed study of k -mer length selection could also include recursive scanning of a reference genome for maximum repeat length and a priori comparison to the genomes of similar organisms. Continued effort to understand and evaluate the decisions used when simplifying or error correcting the de Bruijn graph will lead to higher quality assemblies and serve to unify the field. This includes statistical decision making as well as reference to biological markers and archived genomic data.

References

- [1] Watson, J.D., Francis, H.C.: A Structure for DNA. *Nature* 171, 737–738 (1953)
- [2] Watson, J.D., Francis, H.C.: Genetical Implications of the structure of Deoxyribonucleic Acid. *Nature* 171, 964–967 (1953)
- [3] Gilbert, W., Maxam, A.: The nucleotide seq. of the lac operator. *Proc. Natl. Acad. Sci. U.S.A* 12(70), 3581–3584 (1973)
- [4] Sanger, F., Nicklen, S., Coulson, R.A.: DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A* 12(74), 5463–5467 (1977)
- [5] Tamarin, R.H.: Principles of Genetics, 4th edn. Wm. C. Brown Publishers (1993)
- [6] Ewing, B., Phil, G.: Base-Calling of Automated Sequencer Traces UsingPhred. II. Error Probabilities. *Genome Res.*, 186–194 (1998)
- [7] Ewing, B., et al.: Base-Calling of Automated Sequencer Traces UsingPhred. I. Accuracy Assessment. *Genome Res.* 8, 175–185 (1998)
- [8] Ronaghi, M., Uhlén, M., Nyrén, P.: A sequencing method based on real-time pyrophosphate. *Science* 363, 365 (1998)
- [9] Margulies, M., et al.: Genome Sequencing in Open Microfabricated High Density Picoliter Reactors. *Nature* 437, 376–380 (2003)
- [10] Altschul, S.F., et al.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acid Res.* 25, 3389–3402 (1997)

- [11] Linz, P.: An Intro. to Formal Lang. & Automata, 4th edn. Jones & Bartlett, Boston (2006)
- [12] Pevzner, P.A.: 1-Tuple DNA sequencing: computer analysis. *J. Biomol. Struct. Dyn.* 7, 63–73 (1989)
- [13] Ramana, M.I., Michael, W.S.: A New Algorithm for DNA Sequence Assembly. *Journal of Computational Biology* 2(2), 291–306 (1995)
- [14] Zerbino, D.R., Ewan, B.: Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* 18, 821–829 (2008)
- [15] Gross, J.L., Yellen, J.: *Handbook of graph theory*. CRC Press, Boca Raton (2004); 69 DRAFT 04/02/2010 AE LLC
- [16] Blattner, F.R.: The complete genome sequence of *E. coli* K-12. *Sci.*, 1453–1462 (1997)
- [17] Richter, D.C., et al.: MetaSim—A Sequencing Simulator for Genomics and Metagenomics. *PLoS ONE* 3(1), e3373 (2008)
- [18] Schatz, M.C., et al.: Hawkeye: an interactive visual analytics tool for genome assemblies. *Genome Biology* 8, R34 (2008)
- [19] NCBI. FASTA format description,
<http://www.ncbi.nlm.nih.gov/blast/fasta.shtml>
- [20] Kyoto University Bioinformatics Center. GenomeNet (March 22 (2010),
<http://www.genome.jp>
- [21] Leipzig: Standardized-velvet-assembly-report - Project Hosting on Google Code (March 22, 2010), <http://code.google.com/p/standardized-velvet-assembly-report>
- [22] Roche Diagnostics Co. Products & Solutions - Syetem Benefits: 454 Life Sciences, a Roche Company (March 22, 2010),
<http://454.com/products-solutions/system-benefits.asp>

A Novel Digital Video Watermarking Scheme Based on the Scene Change Analysis

Tummalapalli Geetamma¹, T.V.N.N.M. Vamsi Krishna², and D. Srinivasa Rao²

¹ Dept. of ECE, GMRIT, A.P., India

² Mahindra Satyam, Hyderabad, India

tgeetamma@yahoo.co.in

Abstract. The rapid expansion of the internet in the past few years has enormously increased the availability of digital data such as audio, images and video to the public. This motivates the protection of digital information against illegal duplication and manipulation. Digital watermarking is a technique for protecting the copyright of intellectual property which has become an active research area recently. Watermarking is a process of embedding copyright information or an identification number to the digital data. This paper proposes an oblivious scheme for video watermarking. This proposes the idea of embedding different parts of single watermark into different scenes of a video. The watermark is extracted from the watermarked image without any knowledge of the original image.

Keywords: Discrete Wavelet Transformation, Digital Watermarking, Scene change, Video.

1 Introduction

One of the reasons for the rapid development in research in digital watermarking is the need to find a solution for protecting intellectual properties of digital material. In order to embed watermark information in host data, watermark embedding techniques apply minor modifications to the host data in a perceptually invisible manner, where the modifications are related to the watermark information. The watermark information can be retrieved afterwards from the watermarked data by detecting the presence of these modifications. Most of the watermarking techniques proposed till date are applicable to gray scale images, but can be easily extended to color images by watermarking luminance component. Most of the existing watermarking algorithms can be classified according to the following criteria. The selection of locations where the watermark is embedded. The domain in which algorithm operates. For example, an algorithm can modify the image in the spatial domain directly to embed the watermark or it can transform the image in to other domains like DCT [14], DFT [12], DWT [13] and fractal.

2 Discrete Wavelet Transform (DWT)

2.1 The Fast Wavelet Transform Algorithm

The Discrete Wavelet Transform (DWT) coefficients can be computed by using Mallat’s Fast Wavelet Transform algorithm. This algorithm is sometimes referred to as the *two-channel sub-band coder* and involves filtering the input signal based on the wavelet function used. To explain the implementation of the Fast Wavelet Transform algorithm consider the following equations:

$$\phi(t) = \sum_k c(k)\phi(2t - k) \tag{1}$$

$$\psi(t) = \sum_k (-1)^k c(1 - k)\phi(2t - k) \tag{2}$$

$$\sum_k c_k c_{k-2m} = 2\delta_{0,m} \tag{3}$$

The first equation is known as the twin-scale relation (or the dilation equation) and defines the scaling function Φ . The next equation expresses the wavelet Ψ in terms of the scaling function Φ . The third equation is the condition required for the wavelet to be orthogonal to the scaling function and its translates. The high pass filter is obtained from the low pass filter using the relationship,

$$g_k = (-1)^k c(1 - k) \tag{4}$$

Starting with a discrete input signal vector s , the first stage of the FWT algorithm decomposes the signal into two sets of coefficients. These are the approximation coefficients $cA1$ (low frequency information) and the detail coefficients $cD1$ (high frequency information), as shown in the Fig. 1 below.

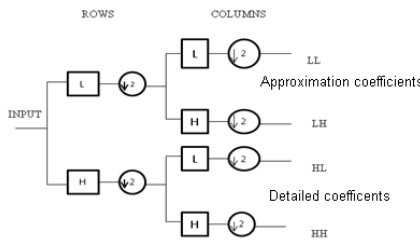


Fig. 1. Filter operation during DWT

The coefficient vectors are obtained by convolving s with the low-pass filter Lo_D for approximation and with the high-pass filter Hi_D for details. This filtering operation is then followed by dyadic decimation or down sampling by a factor of 2.

2.2 Signal Reconstruction

As shown in Fig. 3 the original signal can be reconstructed or synthesized using the inverse discrete wavelet transform (IDWT).The synthesis starts with the approximation

and detail coefficients cA_j and cD_j , and then reconstructs cA_{j-1} by up sampling and filtering with the reconstruction filters.

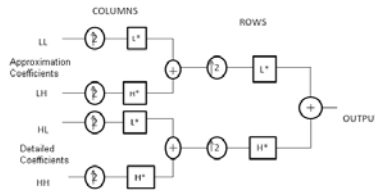


Fig. 2. Wavelets reconstruction

3 Oblivious Video Watermarking Using Scene Change Analysis

3.1 A Proposed Video Watermarking Method

With the rapid growth of the Internet and multimedia systems in distributed environments, it is easier for digital data owners to transfer multimedia documents across the Internet. Therefore, there is an increase in concern over copyright protection of digital contents [1]-[4]. Traditionally, encryption and control access techniques were employed to protect the ownership of media. These techniques, however, do not protect against unauthorized copying after the media have been successfully transmitted and decrypted. Recently, watermark techniques are utilized to maintain the copyright [4]-[7]. Video watermarking introduces a number of issues not present in image watermarking. Due to a large amount of data and inherent redundancies between frames, video signals are highly susceptible to piracy attacks, including frame averaging, frame dropping, frame swapping, statistical analysis, etc. This problem can be overcome by applying scene change detections and scrambled watermarks in a video.

The scheme is robust against frame dropping, as the same part of the watermark is embedded into the frames of a scene. For different scenes, different parts of the watermark are used, making the scheme robust against frame averaging and statistical analysis [8].

3.2 Scene-Based Video Watermarking Scheme

The new watermarking scheme we propose is based on scene changes in [8]. Fig 5 shows an overview of watermarking process. In this scheme, a video is taken as the input, and then a watermark is decomposed into different parts which are embedded in corresponding frames of different scenes in the original video. As applying a fixed image watermark to each frame in the video leads to the problem of maintaining statistical and perceptual invisibility [20] this scheme employs independent watermarks for successive but different scenes. However, applying independent watermarks to each frame also presents a problem if regions in each video frame remain little or no motion frame after frame. These motionless regions may be statistically compared or averaged to remove the independent watermarks [16]. Consequently, an identical

watermark is used within each motionless scene. With these mechanisms, the proposed method is robust against the attacks of frame dropping, averaging, swapping, and statistical analysis. This scheme consists of four parts:

1) *Watermark Preprocess*

A watermark is scrambled into small parts in a preprocess, and they are embedded into different scenes so that the scheme can resist a number of attacks toward to the video. A 256-greylevel image is used as the watermark, so 8 bits can represent each pixel. The watermark is first scaled to a particular size as follows:

$$2^n \leq m; n > 0(5)$$

$$p + q = n; p, q > 0(6)$$

Where m is the number of scene changes and p, q, n are positive integers. The size of the watermark is represented as

$$4.2^p \times 4.2^q(7)$$

To make the scheme more robust, the processed watermarks are transformed to the wavelet domain. Sample preprocessed watermarks are shown in figures below where Fig. 4 is the original watermark, Fig. 6 (w1) – (w8) represent the small watermarks in the spatial domain.



Fig. 3. Original Watermarking

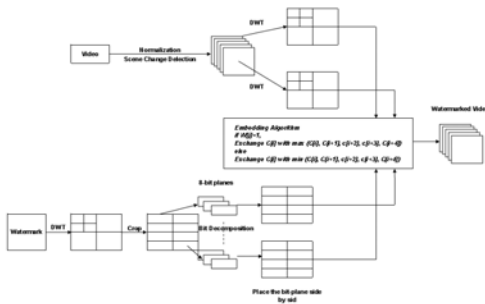


Fig. 5. Overview of watermarking process

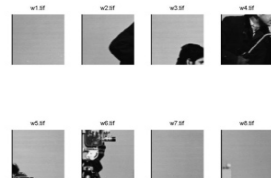


Fig. 6. w1-w8 Preprocessed watermarks

2) Video Preprocess

The watermark scheme is based on 4-level DWT. All frames in the video are transformed to the wavelet domain. The frames are decomposed in 4-level subband frames by separable two-dimensional (2-D) wavelet transform. It produces a low-frequency subband LL, and three series of high-frequency sub bands LH, HL, HH. According to the energy distribution, LL is the most important than LH, HL, HH. For different levels, the higher the level, the more important the sub bands.

In this scheme, only embed the watermark in the middle frequency sub bands this scheme is based on 4-levels DWT, which is determined by experiments. If less than 4-levels is applied, the capacity of the scheme would be decreased; if larger than 4-levels is applied, the quality of the watermark video is affected.

Scene changes are detected from the video by applying the histogram difference method on the video stream. If the difference of the two scenes is greater the threshold, consider there is a scene change. The threshold is again determined by experiments. Independent watermarks are embedded in frames of different scenes. Within a motionless scene, an identical watermark is used for each frame. Watermark is used for the first scene. When there is a scene change, another watermark is used for the next scene. The watermark for each scene can be chosen with a pseudo-random permutation such that only a legitimate watermark detector can reassemble the original watermark.

3) Watermark Embedding Algorithm in Video stream

The watermark is then embedded to the video frames by changing position of some DWT coefficients with the following condition:

```

If  $W_j=1$  then
Exchange  $\max(C_i, C_{i+1}, C_{i+2}, C_{i+3}, C_{i+4})$ 
else
Exchange  $\min(C_i, C_{i+1}, C_{i+2}, C_{i+3}, C_{i+4})$ 
end if

```

where C_i is the i th DWT coefficient of a video frame, and W_j is the j th pixel of a corresponding watermark image. When $W_j = 1$, we perform an exchange of the C_i with the maximum value among $C_i, C_{i+1}, C_{i+2}, C_{i+3}, C_{i+4}$. When $W_j = 0$, we perform an exchange of the C_i with the minimum value among $C_i, C_{i+1}, C_{i+2}, C_{i+3}, C_{i+4}$. With this algorithm, the retrieval of the embedded watermark does not need the original image. The higher frequency coefficients of the watermark are embedded to higher frequency parts of the video frame, and only the middle frequency wavelet coefficient of the frame (middle frequency subband) is watermarked [16].

4) Watermark Extraction Algorithm

The video is processed to detect the video watermark. In this step, scene changes are detected from the tested video. Also, each video frame is transformed to the wavelet domain with 4-levels. Then the watermark is extracted with the following condition: where WC_i is the i th DWT coefficient of a watermarked video frame, and EW_j is the j th pixel of the extracted watermark [17]. When the watermark WC_i is greater

than median value among $WC_i, WC_{i+1}, WC_{i+2}, WC_{i+3}, WC_{i+4}$ the extracted watermark EW_j is considered as one, i.e., 1; otherwise, it is considered as zero, i.e., 0. With this algorithm, the retrieval of the embedded watermark does not need the original image. This is an important property to video watermarking.

```

if
     $WC_i > \text{median}(WC_i, WC_{i+1}, WC_{i+2}, WC_{i+3}, WC_{i+4})$ 
then
     $EW_j = 1$ 
else
     $EW_j = 0$ 
end if
    
```

As an identical watermark is used for all frames within a scene, multiple copies of each part of the watermark may be obtained. The watermark is recovered by averaging the watermarks extracted from different frames. This reduces the effect if the attack is carried out at some designated frames. Thus, we can combine the 8 bit-planes and recover the 16 x 8 size image, i.e., part of the original watermark. If enough scenes are found and all parts of the watermark are collected, the original large watermark image can be reconstructed.

4 Results



Fig. 4. Watermarked sequence

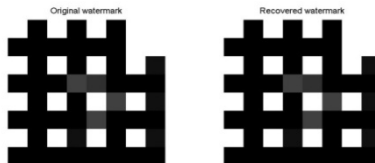


Fig. 6. Watermark used

5 Conclusion

This paper proposes an innovative scene based video watermarking scheme. The process of this comprehensive video watermarking scheme, including watermark preprocessing, video preprocessing, watermark embedding and watermark detection is implemented. Also, this technique is robust against many attacks because of embed watermark in DWT domain. This can be extended with the encryption technique to increase the security of watermarks. This proposed watermark scheme can further be associated with different applications to achieve a sophisticated system and the fidelity can be improved by applying genetic algorithm.

References

1. Piva, A., Bartolini, F., Barni, M.: Managing copyright in open networks. *IEEE Trans. Internet Computing* 6(3), 18–26 (2002)
2. Lu, C., Yuan, H., Liao, M.: Multipurpose watermarking for image authentication and protection. *IEEE Trans. Image Process.* 10(10), 1579–1592 (2001)
3. Lu, C., Huang, S., Sze, C., Liao, H.Y.M.: Cocktail watermarking for digital image protection. *IEEE Trans. Multimedia* 2(6), 209–224 (2000)
4. Lee, J., Jung, S.: A survey of watermarking techniques applied to multimedia. In: *Proc. 2001 IEEE Int. Symp. Industrial Electronics (ISIE)*, vol. 1, pp. 272–277 (2001)
5. Barni, M., Bartolini, F., Caldelli, R., De Rosa, A., Piva, A.: A robust watermarking approach for raw video. Presented at the 10th Int. Packet Video Workshop, Cagliari, Italy, May 1–2 (2000)
6. Petitcolas, F. (ed.): *Information Hiding Techniques for Steganography and Digital Watermarking* Stefan Katzenbeisser. Artech House, Norwood (1999)
7. Eskicioglu, A., Delp, E.: An overview of multimedia content protection in consumer electronics devices. *Proc. Signal Processing Image Communication* 16(2001), 681–699 (2001)
8. Chan, P.-W., Lyu, M.R.: A DWT-Based Digital Video Watermarking Scheme with Error Correcting Code. In: Qing, S., Gollmann, D., Zhou, J. (eds.) *ICICS 2003*. LNCS, vol. 2836, pp. 202–213. Springer, Heidelberg (2003)
9. Memon, N.: Analysis of LSB based image steganography techniques Chandramouli. In: *Proc. 2001 Int. Conf. Image Processing*, vol. 3, pp. 1019–1022 (October 2001)
10. Langelaar, G., Setyawan, I., Lagendijk, R.: Watermarking digital image and video data. *IEEE Signal Process. Mag.* 17(9), 20–43 (2000)
11. Mobasser, B.: Direct sequence watermarking of digital video using m-frames. In: *Proc. 1998 Int. Conf. Image Processing*, vol. 2, pp. 399–403 (October 1998)
12. Pereira, S., Pun, T.: Robust template matching for affine resistant image watermarks. *IEEE Trans. Image Process.* 9(6), 1123–1129 (2000)
13. Hong, I., Kim, I., Han, S.: A blind watermarking technique using wavelet transform. In: *Proc. IEEE Int. Symp. Industrial Electronics*, vol. 3, pp. 1946–1950 (1995)
14. Duan, F., King, I., Xu, L., Chan, L.: Intra-block algorithm for digital watermarking. In: *Proc. IEEE 14th Int. Conf. Pattern Recognition*, vol. 2, pp. 1589–1591 (August 1998)
15. George, M., Chouinard, J., Georganas, N.: Digital watermarking of images and video using direct sequence spread spectrum techniques. In: *Proc. 1999 IEEE Canadian Conf. Electrical and Computer Engineering*, vol. 1, pp. 116–121 (May 1999)

16. Swanson, M., Zhu, B., Tewfik, A.: Multiresolution video watermarking using perceptual models and scene segmentation. In: Proc. Int. Conf. Image Processing, Washington, DC, vol. 2, pp. 558–561 (October 1997)
17. Cox, J., Kilian, J., Leighton, F., Shamoon, T.: Secure spread spectrum watermarking for multimedia. *IEEE Trans. Image Process.* 612, 1973–1987 (1997)
18. Petitcolas, F., Anderson, R.: Evaluation of copyright marking systems. In: Proc. IEEE Multimedia Systems, Florence, Italy, pp. 574–579 (June 1999)
19. Kutter, M., Petitcolas, F.: A fair benchmark for image watermarking systems. In: Proc. Electronic Imaging 1999, Security and Watermarking of Multimedia Contents, vol. 3657, pp. 226–239 (1999)
20. Checcacci, N., Barni, M., Bartolini, F., Basagni, S.: Robust video watermarking for wireless multimedia communications. In: Proc. 2000 IEEE Wireless Communications and Networking Conf., vol. 3, pp. 1530–1535 (2000)

Application of Swarm Intelligence Computation Techniques in PID Controller Tuning: A Review

Soumya Ghosal¹, Rajkumar Darbar², Biswarup Neogi³,
Achintya Das⁴, and Dewaki N. Tibarewala⁵

¹ IT Dept., RCC Institute of Information Technology, Kolkata, India
soumyaghosal.2008@gmail.com

² IT Dept., Durgapur Institute of Advanced Tech and Management, Durgapur, India
rajdarbar.r@gmail.com

³ ECE Dept., Durgapur Institute of Advanced Tech and Management, Durgapur, India
biswarupneogi@gmail.com

⁴ ECE Dept., Kalyani Government Engineering College, India
achintya_das123@yahoo.co.in

⁵ School of BioScience & Engineering, Jadavpur University, India
biomed_ju@yahoo.com

Abstract. Swarm Intelligence Computation technique is one of the recent and advanced research topic in the field of Artificial Intelligence. This nature – inspired, global optimization technique is used rapidly in various fields , specially it has become one of the most useful method for efficiency improvement of control and distributed optimization aspects. A review study on tuning of PID controller with effective and satisfactory performance analysis via different swarm intelligence computation techniques is presented in this paper. Tuning of PID via traditional methods and genetic algorithm and their limitations in proper tuning, different structure of PID controllers with the objectives for PID tuning and an efficient intelligent PID controller design is presented in the beginning of this paper. Then a brief literature review on PID tuning with different Swarm Intelligence(SI) techniques i.e. Ant Colony Optimization(ACO), Particle Swarm Optimization(PSO), and Bacterial Foraging Optimization Algorithm(BFOA) as well as their advantages and disadvantages in proper tuning is presented in the afterwards . And finally a performance comparison with simulation results of PID tuning via ZN, GA, PSO, BFOA are experimented on four set of system transfer functions and are studied for effective analysis.

Keywords: Swarm Intelligence Computation, Particle Swarm Optimization(PSO), Ant Colony Optimization(ACO), Bacterial Foraging Optimization Algorithm(BFOA), PID controller tuning.

1 Introduction

PID control is a generic feedback control technology and it is vastly used in automatic controllers in industrial control systems. The PID control was first introduced in 1939 in the market and has been successfully used as controller in process control until today. The basic function of the controller is to execute an algorithm based on the control engineers input and hence to maintain the output at a level so that there is no difference between the

process variable and the setpoint[1]. The term ‘PID’ is an acronym for “proportional, integral, and derivative.” A PID controller is a controller that includes elements with those three functions. The popularity of such kind of controller is due to their functional simplicity and reliability. They provide robust and reliable performance for most systems and the PID parameters are tuned to ensure a satisfactory closed loop performance [2]. A PID controller improves the transient response of a system by reducing the overshoot, and by shortening the settling time of a system [3]. For this control loop to function properly, the PID loop must be properly tuned. Ziegler-Nichols[4], Cohen-Coons[5], Astrom and Haggglund[6] tuning methods were some of the primitive tuning methods for PID control. But, due to having greater phase lag, greater overshoot and insufficient tuning results a linear empirical formula had to be introduced. Genetic Algorithm(GA) was an effective solution for these problems. GA was one of the evolutionary computation tuning approaches that could produce better results in PID tuning than the primitive tuning methods[7] having stochastic global searching characteristics that could mimic the process of natural evolution. But later, a set of new intelligent approaches unitedly called swarm intelligence tuning, such as Ant Colony Optimization, Particle swarm optimization, Bacterial Foraging optimization algorithm were introduced which could produce an effective characteristics of positive feedback, search mechanism, distributed computation and constructive greedy heuristic for simpler, efficient and faster tuning than primitive and other evolutionary tuning approaches for having improved characteristics like dynamic adjustment inertia weight factors, etc. Positive feedback search produces advantageous results, distributive computation can be used to avoid pre-matured convergence and greedy heuristic is helpful to find the solution of early stages of search process. The brief survey on PID tuning via different SI techniques is presented later in this paper.

2 The Different Structure of PID Controller and Intelligent PID Controller Design

The practical difficulty with PID control technology is a lack of industrial standards, which has resulted in 46 different architectures of PID controller. For example, the classical architecture of PID controllers is given below.

$$G_c(s) = K_c \left(1 + \frac{1}{T_i s}\right) \frac{1 + sT_d}{1 + \frac{sT_d}{N}}$$

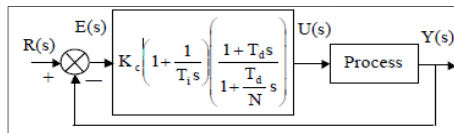


Fig. 1. Classical PID controller in a unity feedback block diagram representation

PID controllers are used in more than 95% of closed-loop industrial processes. Mostly the interest lies in four major characteristics of the closed-loop step response. They are rise time, overshoot, settling time, steady-state error. How the increment of the PID parameters (Kp, Ki, Kd) value can affect the system dynamics is presented in a tabular format in the following way:

Table 1. Performance requirements and objective of PID tuning

Response	Rise Time	Overshoot	Settling Time	S-S Error
K_p	Decrease	Increase	NT	Decrease
K_i	Decrease	Increase	Increase	Eliminate
K_d	NT	Decrease	Decrease	NT

NT: No definite trend Minor change.

In the next figure, the block diagram of an intelligent PID Controller is given.

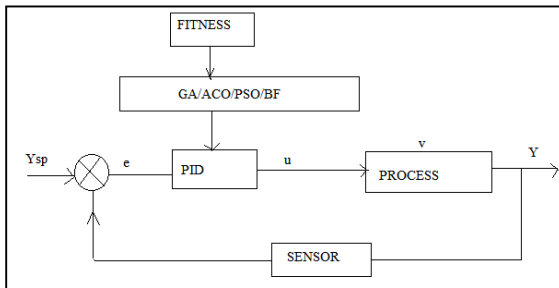


Fig. 2. Block Diagram of an intelligent PID controller[8]

In PID controller, the proportional value determines the reaction of the current error, the integral value determines the reaction based on the sum of recent errors, and derivative value determines the reaction based on the rate at which the error has been changing the weighted sum of these three actions is used to adjust the process via the final control element.

3 Survey on Swarm Intelligence Computation

Swarm intelligence is an algorithm or a device, which is designed for solving distributed problems. It was illuminated by the social behavior of gregarious insects and other animals [9]. Compared with the grads algorithm and traditional evolutionary computations, swarm intelligence has following advantageous characteristics: (a) The cooperating particle of the swarm is distributed; (b) There is no control and data of the center and the system is more robust; (c) It can realize cooperation with indirect communication instead of direct communication and the system is more easily to extend; (d) The ability of particle in the population is simple, the operating time of every particle is also very short and it is easy to be realized. Different swarm intelligence algorithms and their uses in PID tuning are discussed below.

4 Ant Colony Optimization(ACO) in PID Tuning

ACO's are very much suited for finding solutions to different optimization problems. A colony of artificial ants cooperates to find good solutions, which are an emergent property

of the ant's cooperative interaction .Based on their similarities with ant colonies in nature, ant algorithms are adaptive and robust and can be applied to different versions of the same problem as well as to different optimization problems. The main idea behind ant colony optimization is that when the ants search for food, they initially explore the area surrounding their nest randomly. When one finds a food source, it evaluates it, take some food and goes back to the nest. As they move back, they deposit on the ground a chemical substance called pheromone, which is detectable by other ants. The amount of pheromone that is deposited varies depending on the quantity and quality of the food, and leads other ants to that food source. By the use of this property, the ants can find the shortest path between their nest and the source [11]. A basic algorithm for ACO is given in figure 3.

ACO can be used in PID tuning in the following steps-i) Initialize no. of ants, ii) Then run the process model, iii) Evaluate fitness function, iv) Then update the probability, v) After that calculate the optimum of Kp,Ki,Kd values and finally after reaching the maximum iteration the process stops. ACO was successfully used for PID controller tuning from time as it was able to calculate the optimum value of PID parameters in a very effective manner. A literature review on PID tuning via ACO is given below.

```

input: An instance  $P$  of a CO problem model  $P = (S, f, \Omega)$ .
InitializePheromoneValues( $T$ )
 $s_{bs} \leftarrow \text{NULL}$ 
while termination conditions not met do
     $\mathcal{S}_{iter} \leftarrow \emptyset$ 
    for  $j = 1, \dots, n_a$  do
         $s \leftarrow \text{ConstructSolution}(T)$ 
        if  $s$  is a valid solution then
             $s \leftarrow \text{LocalSearch}(s)$  {optional}
            if  $(f(s) < f(s_{bs}))$  or  $(s_{bs} = \text{NULL})$  then  $s_{bs} \leftarrow s$ 
             $\mathcal{S}_{iter} \leftarrow \mathcal{S}_{iter} \cup \{s\}$ 
        end if
    end for
    ApplyPheromoneUpdate( $T, \mathcal{S}_{iter}, s_{bs}$ )
end while
output: The best-so-far solution  $s_{bs}$ 
    
```

Fig. 3. Algorithm of ACO[10]

A research work based on Ant Colony Search-PID tuning is briefly discussed in this paper[51]. In this paper the generation of nodes and path step, the PID controller parameters Kp, Ki, and Kd as the optimized variables, and assumed that each of them has five valid digits, one digit before decimal point and four digits after decimal point and the values of Kp, Ki, and Kd were put on plane O-XY. Here, each decimal digit represents node and connection between them represents the moving path of ants. In transition step, an ant selects the following transition rule :

$$j = \underset{u \in J^k_i}{\operatorname{argmax}} \left\{ \left| \tau(x_i, y_{iu}) \right| \cdot \left| \eta(x_i, y_{iu}) \right|^\beta \right\}, \text{ if } q \leq q_0$$

$$J = J, \text{ if } q > q_0 \tag{1}$$

In formula (1), τ was considered as the pheromone concentration and $\eta(x_i, y_{ij})$ was the visibility of node (x_i, y_{ij}) and it was proposed as:

$$\eta(x_i, y_{ij}) = \frac{10 - |y_{ij} - y_{ij}^*|}{10} \tag{2}$$

When all of the ants in the colony complete their tours once in the ACS-PID algorithm, the pheromone concentration of each node belonging to the best tour, since the beginning of the trial is updated by the following formulas:

$$\begin{aligned} \tau(x_i, y_{ij}) &\leftarrow (1-\rho) \cdot \tau(x_i, y_{ij}) + \rho \cdot \Delta\tau(x_i, y_{ij}) \\ \Delta\tau(x_i, y_{ij}) &= \frac{Q}{ITAE^*} \end{aligned} \tag{3}$$

Here, ρ is the parameter which governs the pheromone decay. The local update is performed as follows:

$$\tau_{ij}(x_i, y_{ij}) \leftarrow (1-\rho) \cdot \tau_{ij}(x_i, y_{ij}) + \rho \cdot \tau_0 \tag{4}$$

PID tuning using this adaptive ACS method produces better results than GA-PID, SA-PID, DE-PID. Another approach taken on PID controller tuning via ACO was based on creating incrementally the construction of solutions based on a probabilistic choice of the solution components [12]. In this paper ACO was used for continuous domains, and it was accomplished by the use of a PDF. A PDF could be represented as any function $P(x) > 0$, such that x that could meet the requirement,

$$\int_{-\infty}^{\infty} P(x)dx = 1 \tag{5}$$

Based on the decision variables $X_i, i = 1, 2, \dots, n$, each ant constructed a solution performing n steps. At an iteration i , an ant was set to choose a value for the variable X_i . After this, a Gaussian kernel was created for this iteration. This Gaussian kernel was denoted by this equation:

$$G^i(x) = \sum_{l=1}^k \omega_l g_l^i(x) = \sum_{l=1}^k \frac{1}{\sigma_l^i \sqrt{2\Pi}} e^{-\frac{(x-\mu_l^i)^L}{2\sigma_l^{2i}}} , i=1,2,\dots,n. \tag{6}$$

where μ was the mean, ω was the weight, and σ was the standard deviation.

The second step of the algorithm was the pheromone update. All the pheromone information was stored in a solution archive T . For each solution to a problem of n dimensions, the algorithm stored in T the values of its n variables, besides the value of the objective function $f(s)$. All the solutions on this archive was evaluated and then ranked. By this rank they could be sorted. The third and last step was just to update the best solution found, in order that it could be shown when the stop conditions met. A modified ant colony optimization incorporating differential evolution was used in this work. In this design the mutation operation was generated by this function:

$$X_i(t+1) = X_{best}(t) + MF \cdot [X_{i_2}(t) - X_{i_3}(t)] \tag{7}$$

In the above equations, t is the time (generation), $MF > 0$ is a real parameter, called mutation factor, which controlled the amplification of the difference between two individuals so as to avoid search stagnation. The mutation operation selected the best

vector $X_{best}(t)$. Then, two individuals were randomly selected and the difference vector was calculated. From the result, it was obtained that with the same preset maximum number of generations, Modified ACO obtained better mean F and minimum F than GA, ES, and ACO methods. Thus, ACO and Modified ACO proved superior in PID tuning than ES and GA. In another work[13] Lyapunov function was proposed for accomplishing robust global convergence for tracking error. Parameter tuning of PID was used by Grid-based searching adopted ACO and self adaptive control strategy was also adopted for pheromone decay. Another PID tuning procedure via ACO was furnished by position tracing and elitist strategy was adopted via improved ACO proposed in work[14]. Important ACO optimization, convergence and application strategies are surveyed and published in this journal paper[15].

A research on satisfying the real time control and obtaining better performance, application of ACA (Ant Colony Algorithm) to optimize the parameters of NN-PID controller to improve the on-line self-tuning capability of this controller is presented in this paper[16]. Tuning of Fuzzy PID controller via ACO is also another research work in PID tuning via SI[17]. This paper proposes a proper framework of PID tuning via proper implementation and simulation approach via ACO. One of the nature based algorithm named artificial bee colony similar to ACO was used in PID tuning[18], which was presented by Karaboga in 2005 to optimize numeric benchmark functions [19]. Trajectory tracking comparison for GA and ACO for PID tuning is presented in this paper[20].

ACO is very much efficient in discrete control optimization. It can produce results where the source and destination is known and predetermined. For, crisp result ACO is productive, though it produces some disadvantages also. Its theoretical analysis is difficult for ACO as well as the research on it is experimental rather than theoretical. Also, the time of convergence is uncertain for ACO. PSO is a simpler version of optimization technique for users and also a developed version of this simulation based approach produces better performances in dynamic optimization and constraint handling. In case of problems that are fuzzy in nature usage of PSO is beneficial rather than ACO.

5 Particle Swarm Optimization and Bacterial Foraging Optimization Algorithm in PID Tuning

Particle Swarm Optimization (PSO) is one of the optimization and kind of evolutionary computation technique. The technique is derived from research on swarm such as bird flocking and fish schooling. In the PSO algorithm, instead of using evolutionary operators such as mutation and crossover to manipulate algorithms, a flock of particles are put into the d-dimensional search space with randomly chosen velocities and positions knowing their best values. The velocity and position of each particle, adjusted accordingly to its own flying experience and the other particles flying experience [9]. It was first introduced by Eberhart and Kennedy in 1995[21]. The description of PSO algorithm is given below [22] where the equations are given as (8) and (9) for iteration via PSO.

$$v_{ij}^{(k+1)} = w \times v_{ij}^{(k)} + c_1 \times \text{rand}() \times (pbest_{ij} - x_{ij}^{(k)}) + c_2 \times \text{rand}() \times (gbest_j - x_{ij}^{(k)}) \quad (8)$$

$$x_{ij}^{(k+1)} = x_{ij}^{(k)} + v_{ij}^{(k+1)} \quad (9)$$

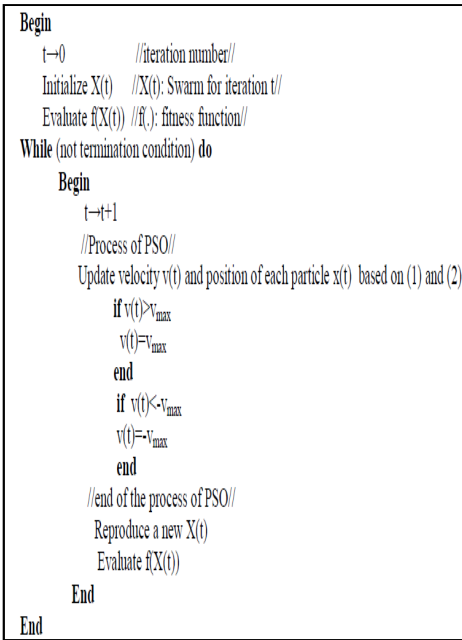


Fig. 4. Algorithm of PSO

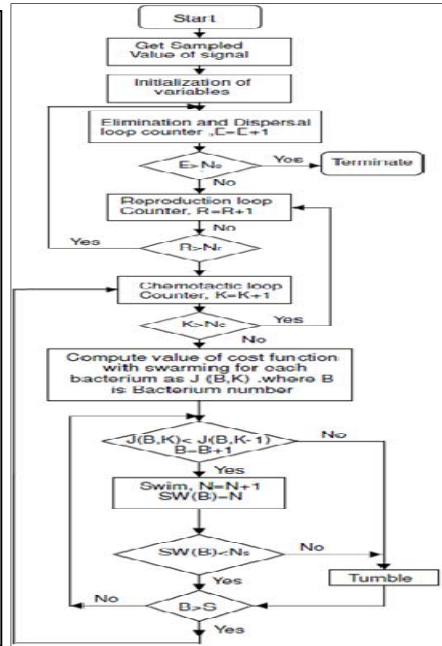


Fig. 5. Flowchart of BFOA

From last decade, Particle Swarm Optimization has been successfully applied in the field of control, design, telecommunication and combinatorial optimization procedures. Around three hundred PSO algorithms are exploited till date. A noble review was done on PSO by De Falco et al[23]. Although PSO has been used mainly to solve unconstrained, single-objective optimization problems, PSO algorithms have been developed to solve constrained problems, multi-objective optimization problems, problems with dynamically changing landscapes, and to find multiple solutions. The application of PSO in nonlinear control domain is also huge. A literature review on its applications in PID controller tuning is discussed below. A research work on designing a robust H₂/H_∞ PID controller design via PSO in shown in[21]. The controller was designed such that the nominal closed loop system was asymptotically stable and robust stability satisfied the required inequality equation. And the disturbance attenuation performance satisfied the following inequality.

$$J_b = \|W_2(s).S(s)\|_{\infty} < 1 \tag{10}$$

A balanced performance criterion to minimize both J_a and J_b simultaneously is to minimize J_{∞} , The minimization of tracking error J_2 by taking integration of error $e(t)$ and by taking inverse laplace transformation ,the robust hybrid controller design was established. Three types of performance estimation was done on PID i.e. the integrated absolute error (IAE), or the integral of squared error (ISE), or integral of time absolute error (ITAE) but after comparing the simulation results of PSO optimization it was seen that the ISE was the best to use with the disturbance condition to get robustness. An Optimal Fractional Order Controller for an AVR System using Particle Swarm

Optimization Algorithm was proposed in this paper[24] where the FOPID PSO designing was done in these steps- i) Randomly initializing the individuals of the population including searching points and velocities in the feasible range, ii) For each initial individual k of the population, the values of the performance criterion were calculated, iii) Then comparing each individual's evaluation value with its personal best p_{id} . The best evaluation value among the p_{id} was denoted as p_g . iv) Modifying the member velocity of each individual, v) If the number of iterations reaches the maximum, go to Step vii, otherwise go to Step ii. vii) The latest p_g was the optimal controller parameter. A PID controller tuning via PSO for power system stabilizing is shown in this paper[25]. Though ZN method was also applied for PID tuning, PSO based tuning proved faster results for reaching steady state condition and proficient outcomes for damping of the multimachine power system transient and dynamic disturbance. A fuzzy PID controller tuning via an intelligent PSO and Genetic Algorithm approach for AVR system is depicted in [26]. The Crazyness based PSO used in this work proved superior in tuning than the binary coded GA used with respect to optimal transient performance and lesser computational time. The work of Fuzzy Logic was to extrapolate intelligently and linearly, the nominal optimal gains in order to determine off-nominal optimal gains for on line off nominal system parameters.

Robust PID controller tuning via PSO [27], PID tuning via Advanced PSO by changing few mathematical structure of original PSO[28], a novel approach based on Stochastic Particle Swarm optimization (SPSO), with dominant eigenvalue shift for designing robust decentralized load frequency control system for interconnected power system[29], PID controller tuning for Hybrid PV-FC-Diesel-Battery Micro Grid Scheme for Village/Resort Electricity Utilization via PSO[30] are some of the PSO application in PID tuning domain. PSO applications in PID tuning for AVR system[31], PID tuning in slider-crank mechanism system [31], evolutionary robotic-vision system and tracking[33] are also some of the advanced research works in the field of PID tuning via PSO. As well as speed control of the Brushless DC Motor (BLDCM) servo system[34], the values of the parameters of a proposed fuzzy PID controllers optimization with minimization of sum square error (SSE)[35], are some of the recent PSO application work in PID controller tuning.

PSO is capable of generating these qualities for its use- its intelligent application in various scientific and research fields, no overlapping and mutation calculation, simple in nature, its a real number code and it is decided directly by the solution. Although, it has few drawbacks also- The method easily suffers from the partial optimism, which causes the less exact at the regulation of its speed and the direction, it is unable to work out the problems of scattering and optimization. Also PSO cannot work out the problems of non-coordinate system, such as the solution to the energy field and the moving rules of the particles in the energy fields.

Another important swarm intelligence technique known as Bacterial Foraging Optimization Algorithm(BFOA) was first introduced and developed by Passino[36]. Inspired by social foraging of bacteria E.Coli ,which can be explained by four processes namely Chemotaxis, Swarming, Reproduction, Elimination and Dispersal [36], this nature based algorithm was generated and it has made a good impact as a global optimization algorithm in distributed optimization and control. Its application in science and engineering domain and analysis is very exquisitely discussed in [37]. Fig.5 shows a flow chart of BFOA method. A brief literature review on PID controller tuning via BFOA is given below.

Bacterial Foraging Optimization method is newer than other swarm intelligence methods (PSO,ACO) and still hard research work is going on development of this heuristic method. For its real world application this method is gaining more popularity to researchers. An application of hybrid genetic algorithm and BFOA in global optimization is shown in[38]. It was shown using on PID of AVR systems and simulation results showed the superiority of BFOA in tuning than the GA. Interface suppression of linear antenna array by amplitude control via BFOA is depicted in[39]. It was found that the nulling method based on BFA was capable of steering the array nulls precisely to the undesired interference directions. Designing a bow-tie antenna for 2.45 GHz Radio Frequency Identification (RFID) readers via bacteria swarm optimization method(BSO) and Nelder-Mead algorithm is proposed in[40]. The BSO proved more efficient result results in parameter tuning than convenient PSO in the simulation result. Different applications and methodologies for tuning of PID via this hybrid SI technique can be found out in [41]. A comparative study of BFOA and PSO and state of art version of PSO for optimizing multi-modal and high dimensional functions clearly shows its efficiency in such optimization problem[42]. Its generates fruitful results in its application in neural networks in load forecasting, fuzzy logic based problems, signal processing, pattern recognition, robust bus architecture design for power systems, social scheduling problems, as well as in controller tuning approaches. Its application in PID tuning are presented in these works[43,44,45]. A fuzzy PID controller tuning approach is shown in[46].Implementation of BFOA in PID tuning for SIMO process on FPGA[47], design approach of PID controller with rejection function against external disturbance in motor control system via BFOA[48], PID tuning in power system stability and in AVR system for obtaining minimum errors and to damn optimally by BFOA[49] are some of the recent research applications of BFOA in PID tuning.

Although having such capabilities for using in multiple engineering and science research domains, BFOA method still needs more development and adaptability for making it , not just a successful, but one of the best swarm optimization methods in real world application. Our review work cannot be established without the favorable help of this book[50].

6 Comparative Analysis of PID Tuning via Traditional, GA and Different SI Techniques

In this work, four systems were selected for simulation using various tuning method such as ZN, PSO, GA, BFO. MATLAB 7.0 software is employed. The four set of systems, considered for the tuning comparisons are as follows:

Set 1:

$$G(s) = \frac{1}{0.21s^2 + 0.4038s + 0.0411} \quad (11)$$

Set 2:

$$G(s) = \frac{1}{0.21s^2 + 0.1193s + 0.0245} \quad (12)$$

Set 3:

$$G(s) = \frac{1}{0.21s^2 + 0.0434s + 0.0299} \tag{13}$$

Set 4:

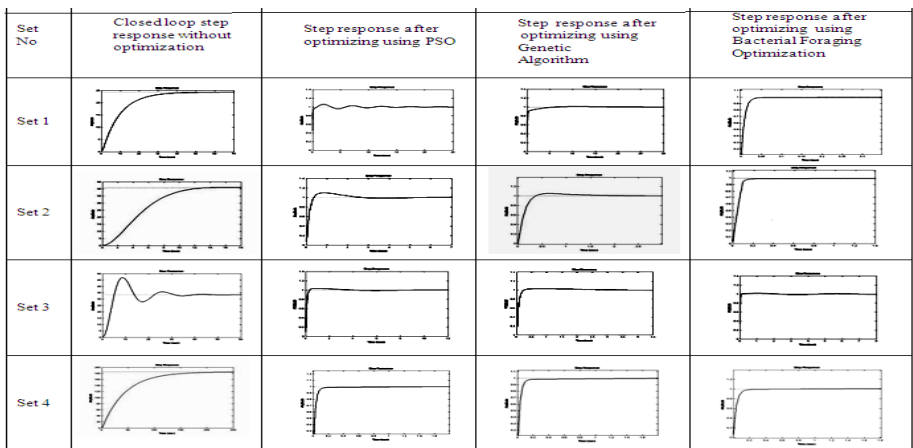
$$G(s) = \frac{1}{0.21s^2 + 0.2369s + 0.0054} \tag{14}$$

The PID parameter values obtained using ZN,GA,PSO and BFO and the closed loop step responses for tuning is shown in next two tables.

Table 2. Different values of PID Parameters after tuning by various tuning methods

Set No	Values of PID parameters using Z-N	Values of PID parameters using PSO	Values of PID parameters using Genetic Algorithm	Values of PID parameters using Bacteria Foraging
Set 1	$K_p=0.0471$ $T_d=1.046$ $T_i=4.161$	$K_p=1.8255$ $T_d=2.816$ $T_i=0.2219$	$K_p=1.2365$ $T_d=2.529$ $T_i=5.859$	$K_p=9.8274$ $T_d=2.826$ $T_i=0.2196$
Set 2	$K_p=0.0083$ $T_d=1.8383$ $T_i=7.3535$	$K_p=1.9045$ $T_d=0.8013$ $T_i=1.0019$	$K_p=2.1023$ $T_d=0.7929$ $T_i=6.9427$	$K_p=1.983$ $T_d=1.7938$ $T_i=6.9938$
Set 3	$K_p=0.0028$ $T_d=2.165$ $T_i=8.66$	$K_p=1.255$ $T_d=1.8728$ $T_i=1.215$	$K_p=1.8937$ $T_d=1.7856$ $T_i=1.0105$	$K_p=4.199$ $T_d=2.8665$ $T_i=0.189$
Set 4	$K_p=0.0036$ $T_d=3.37$ $T_i=13.48$	$K_p=4.478$ $T_d=1.821$ $T_i=3.545$	$K_p=3.146$ $T_d=2.278$ $T_i=4.335$	$K_p=5.0257$ $T_d=1.147$ $T_i=2.726$

Table 3. Comparative diagrams of closed loop step responses for various optimization techniques



For setting up perfections and better efficacies, each set of transfer functions were tuned by different tuning method in this work. Though in comparison, the swarm intelligence methods proved superior than the traditional ZN method which can be inferred from table no 3. From that table, BFOA method is giving the best optimization among all the tuning methods. PSO and GA showed good results than ZN but surely BFOA produces the better tuning than these EA methods.

7 Conclusion

In this paper we have presented a brief literature review of various applications of swarm computation techniques in PID controller tuning research domain. Alongwith a comparative analysis for different EA and Swarm Intelligence techniques for PID tuning is presented as the critical discussions for using those techniques were also mentioned. For, future works, the dynamic determination of best destination for ACO, adapting fitness sharing for PSO, updating velocity for each individual by taking the best element found in all iterations rather than that of current iteration can be taken for proper implementation to make these techniques more effective in PID tuning as well as global optimization which may become a very effective aspect of artificial and computational intelligence in future.

References

- [1] Araki, M.: Control systems. In: Robotics and Automation-Vol II – PID Control. Kyoto University, Japan
- [2] Hwa, K.D., Park, J.: Intelligent PID Controller Tuning of AVR System Using GA and PSO. In: Huang, D.-S., Zhang, X.-P., Huang, G.-B. (eds.) ICIC 2005, Part II. LNCS, vol. 3645, pp. 366–375. Springer, Heidelberg (2005)
- [3] Astrom, K.J., Hagglund, T.: The future of PID control. *Control Eng. Pract.* 9(11), 1163–1175 (2001)
- [4] Ziegler, G., Nichols, N.B.: Optimum settings for automatic controllers. *Trans. ASME* 64, 759–768 (1942)
- [5] Cohen, G.H., Coon, G.A.: Theoretical Consideration of Retarded Control. *Trans. ASME* 75, 827/834 (1953)
- [6] Astrom, K.J., Hagglund, T.: Automatic tuning of simple regulators with specifications on phase and amplitude margins. *Automatica* 20, 645–651 (1984)
- [7] Krohling, R.A., Rey, J.P.: Design of optimal disturbance rejection PID controllers using genetic algorithm. *IEEE Trans. Evol. Comput.* 5(1), 78–82 (2001)
- [8] Nasri, M., Nezamabadi-pour, H., Maghfoori, M.: A PSO-Based Optimum Design of PIO Controller for a Linear Brushless DC Motor. *World Academy of Science, Engineering and Technology* 26 (2007)
- [9] Kennedy, J., Eberhart, R.C.: *Swarm intelligence*. Morgan Kaufmann Publisher, San Francisco (2001)
- [10] Dorigo, M., Blum, C.: Ant colony optimization theory: A survey. *Theoretical Computer Science* 344, 243–278 (2005)
- [11] Socha, K., Dorigo, M.: Ant colony optimization for continuous domains. *European Journal of Operational Research* 185(3), 1155–1173 (2009)
- [12] Coelho, L., Bernert, D.: A modified ant colony optimization algorithm based on differential evolution for chaotic synchronization. *Expert Systems with Applications* 37, 4198–4203 (2010)

- [13] Duan, H., Liu, S., Wang, D., Yu, X.: Design and realization of hybrid ACO-based PID and LuGre friction compensation controller for three degree-of-freedom high precision flight simulator. *Simulation Modelling Practice and Theory* 17(6), 1160–1169 (2009)
- [14] Duan, H., Wang, D., Yu, X.: Novel Approach to Nonlinear PID Parameter Optimization Using Ant Colony Optimization Algorithm. *Journal of Bionic Engineering* 3(2), 73–78 (2006)
- [15] Dorigo, M., Stützle, T.: The Ant Colony Optimization Metaheuristic: Algorithms, Applications, and Advances. In: *Handbook of Metaheuristics*, vol. 57, pp. 250–285. Springer, Heidelberg (2003)
- [16] Cao, C., Guo, X., Liu, Y.: Research on Ant Colony Neural Network PID Controller and Application. In: *Proc. SPND*, pp. 253–258. IEEE, Qingdao (2007)
- [17] Boubertakh, H., Tadjine, M., Glorennec, P.-Y., Labiod, S.: Tuning fuzzy PID controllers using ant colony optimization. In: *Proc. Med.*, pp. 13–18. IEEE, Thessaloniki (2009)
- [18] Abachizadeh, M., Yazdi, M.R.H., Yousefi-Koma, A.: Optimal tuning of PID controllers using Artificial Bee Colony algorithm. In: *AIM 2010*, pp. 379–384. IEEE/ASME, Montreal (2010)
- [19] Karaboga, D.: An idea based on honey bee swarm for numerical optimization: Technical report. Erciyes University, Turkey (2005)
- [20] Ünal, M., Erdal, H., Topuz, V.: Trajectory tracking performance comparison between genetic algorithm and ant colony optimization for PID controller tuning on pressure process. In: *Computer Applications in Engineering Education*. Wiley (2010)
- [21] Zhang, L.: Simplex method based optimal design of PID controller. *Information and Control* 33(3), 376–379 (2004)
- [22] Ali Al-Waily, R.S.: Design of Robust Mixed H₂/H_∞ PID Controller Using Particle Swarm Optimization. *International Journal of Advancements in Computing Technology* 2(5), 53–60 (2010)
- [23] De Falco, Cioppa, A.D., Tarantino: Evaluation of Particle Swarm Optimization Effectiveness in Classification
- [24] Zamani, M., Ghartemani, M., Sadati, N., Parniani, M.: Design of a fractional order PID controller for an AVR using particle swarm optimization. *Control Engineering Practice* 17(12), 1380–1387 (2009)
- [25] Oonsivilai, A., Marungsri, B.: Stability Enhancement for Multi-machine Power System by Optimal PID Tuning of Power System Stabilizer using Particle Swarm Optimization. *WSEAS Transactions on Power Systems* 6(3), 465–474 (2008)
- [26] Mukherjee, V., Ghoshal, S.P.: Intelligent particle swarm optimized fuzzy PID controller for AVR system. *Electric Power Systems Research* 77(12), 1689–1698 (2007)
- [27] Kim, T.H., Maruta, I., Sugie, T.: Particle Swarm Optimization based Robust PID Controller Tuning Scheme. In: *Proc. IEEE Conference on Decision and Control*, New Orleans, LA, USA, pp. 200–205 (2007)
- [28] Jalivand, A., Kimiyaghalam, A., Ashouri, A., Mahdavi, M.: Advanced Particle Swarm Optimization-Based PID Controller Parameters Tuning. In: *Proc. 12th IEEE International Multitopic Conference*, pp. 429–435 (2008)
- [29] Ebrahim, M.A., Mostafa, H.E., Gawish, S.A., Bendary, F.M.: Design of Decentralized Load Frequency Based-PID Controller Using Stochastic Particle Swarm Optimization Technique. In: *Proc. IEEE. EPECS 2009*, Sharjah, pp. 1–6 (2009)
- [30] Sharaf, A., El-Gammal, A.: A novel efficient PSO-self regulating PID controller for hybrid PV-FC-diesel-battery micro grid scheme for village/resort electricity utilization. In: *Proc. IEEE EPEC 2010*, Halifax, NS, pp. 1–6 (2010)
- [31] Gaing, Z.L.: A particle swarm optimization approach for optimum design of PID controller in AVR system. *IEEE Trans. Energy Conversion* 19(2), 384–391 (2004)

- [32] Kao, C.C., Chuang, C.W., Fung, R.: The self-tuning PID control in a slider–crank mechanism system by applying particle swarm optimization approach. *Mechatronics* 16(8), 513–522 (2006)
- [33] Sulistijono, I.A., Kubota, N.: Evolutionary robot vision and particle swarm optimization for Multiple human heads tracking of a partner robot. In: Proc. CEC 2007, pp. 1537–1541. IEEE, Singapore (2007)
- [34] Ren, Y., Xu, X.: Optimization Research of PSO-PID Algorithm for the Design of Brushless Permanent Magnet Machines. In: Fifth IEEE International Symposium on Embedded Computing, Washington, DC, USA (2008)
- [35] Dorrah, H.T., El-Garhy, A.M., El-Shimy, M.E.: Design of PSO-Based Optimal Fuzzy PID Controllers for the Two-Coupled Distillation Column Process. In: Proc. 14th International Middle East Power Systems Conference, pp. 181–188. Cairo University, Egypt (2010)
- [36] Passino, K.M.: Biomimicry of bacterial foraging for distributed optimization and control. *IEEE Control Systems* 22(3), 52–67 (2002)
- [37] Das, S., Biswas, A., Dasgupta, S., Abraham, A.: Bacterial Foraging Optimization Algorithm: Theoretical Foundations, Analysis, and Applications. In: Abraham, A., Hassanien, A.-E., Siarry, P., Engelbrecht, A. (eds.) *Foundations of Computational Intelligence Volume 3*. SCI, vol. 203, pp. 23–55. Springer, Heidelberg (2009)
- [38] Kim, D., Abraham, A., Cho, J.: A hybrid genetic algorithm and bacterial foraging approach for global optimization. *Information Sciences* 177(18), 3918–3937 (2007)
- [39] Guney, K., Busbug, S.: Interference Suppression Of Linearantenna Arrays By Amplitude-Only Control Using A Bacterial Foraging Algorithm. *Progress In Electromagnetics Research, PIER* 79, 475–497 (2008)
- [40] Mahmoud, K.R.: Design Optimization Of A Bow-Tie Antenna For 2.45Ghz Rfid Readers Using A Hybrid BSO-NM Algorithm. *Progress In Electromagnetics Research, PIER* 100, 105–117 (2010)
- [41] Grosan, C., Abraham, A.: Hybrid Evolutionary Algorithms: Methodologies, Architectures, and Reviews. In: *Hybrid Evolutionary Algorithms*, vol. 75, Springer, Heidelberg (2007), doi:10.1007/978-3-540-73297-6_1
- [42] Biswas, A., Dasgupta, S., Das, S., Abraham, A.: Synergy of PSO and Bacterial Foraging Optimization – A Comparative Study on Numerical Benchmarks. In: *Innovations in Hybrid Intelligent Systems*, pp. 255–263. Springer, Heidelberg (2007)
- [43] Kim, D.H., Cho, J.H.: Adaptive Tuning of PID Controller for Multivariable System Using Bacterial Foraging Based Optimization. In: Szczepaniak, P.S., Kacprzyk, J., Niewiadomski, A. (eds.) *AWIC 2005. LNCS (LNAI)*, vol. 3528, pp. 920–928. Springer, Heidelberg (2005)
- [44] Korani, W.M., Dorrah, H.T., Emara, H.M.: Bacterial foraging oriented by Particle Swarm Optimization strategy for PID tuning. In: Proc. CIRA 2009, pp. 445–450. IEEE, Daejeon (2009)
- [45] Oyekan, J., Hu, H.: A Novel Bacterial Foraging Algorithm for Automated tuning of PID controllers of UAVs. In: *Proceedings of the 2010 IEEE International Conference on Information and Automation, China*, pp. 693–698 (2010)
- [46] Su, T.-J., Chen, G.-Y., Cheng, J.-C., Yu, C.-J.: Fuzzy PID controller design using synchronous bacterial foraging optimization. In: Proc. 3rd International Conference on Information Sciences and Interaction Sciences (ICIS), pp. 639–642. IEEE, China (2010)
- [47] Jain, T., Patel, V., Nigam, M.: Implementation of PID Controlled SIMO Process on FPGA Using Bacterial Foraging for Optimal Performance. *IJCEE* 1(2), 107–110 (2009)
- [48] Kim, D.H., Cho, J.H.: Robust Tuning of PID Controller With Disturbance Rejection Using Bacterial Foraging Based Optimization. *WSEAS Transactions on Systems* (2004)

- [49] Manuaba, I., Abdillah, M., Soeprijanto, A., Hery, M.P.: Coordination of PID based power system stabilizer and AVR using combination bacterial foraging technique — Particle swarm optimization. In: Proc. ICMSAO 2011, pp. 1–7. IEEE, Kuala Lumpur (2011)
- [50] Abraham, A., Das, S., Roy, S.: Swarm Intelligence Algorithms for Data Clustering. Soft Computing For Knowledge Discovery And Data Mining Book, Part IV
- [51] Tan, G., Zeng, Q., Li, W.: Design of PID controller with incomplete derivation based on ant system algorithm. *Journal Of Control Theory And Applications* 2(3), 246–252 (2004)

A Steganographic Scheme for Color Image Authentication Using Z-Transform (SSCIAZ)

Nabin Ghoshal¹, Soumit Chowdhury², and Jyotsna Kumar Mandal³

¹ Dept. of Engineering and Technology Studies,
University of Kalyani, Kalyani, Nadia-741235, West Bengal, India

² Dept. of Computer Science and Engineering,
Govt. College of Engineering & Ceramic Technology,
73, A. C. Banerjee Lane, Beliaghata, Kolkata-700010

³ Dept. of Computer Science and Engineering,
University of Kalyani, Kalyani, Nadia-741235, West Bengal, India
{nabin_ghoshal, joy_pinu}@yahoo.co.in,
jkm.cse@gmail.com

Abstract. This paper deals with a novel Steganographic technique which demonstrates the color image authentication in Z-domain based on the Discrete two dimensional Z-Transform. The Transform is applied on mask of sub-image block of size 2×2 of spatial components in row major order for the entire image. Single bit from the authenticating secret message/image is fabricated into the real part of the frequency component of each carrier image byte. A delicate re-adjust phase is incorporated in all components of each mask after embedding, to keep the pixel values positive and non-fractional in the spatial domain. Robustness is achieved through embedding bits in variable positions of carrier image determined by a cyclic Fibonacci series. Experimental results show the enhanced performance of the proposed watermarking technique.

1 Introduction

Copyright [4, 5] abuse is the motivating factor in developing new encryption technologies. One such technology is digital watermarking [8, 9]. One of the driving forces behind the increased use of copyright marking is the growth of the Internet which has allowed images, audio, video, etc to become available in digital [2, 3] form. Though this provides an additional way to distribute material to consumers it has also made it far easier for copies of copyrighted material to be made and distributed. Using the Internet a copy stored on a computer can be shared easily with anybody regardless of distance often via a peer-to-peer network which doesn't require the material to be stored on a server and therefore makes it harder for the copyright owner to locate and prosecute offending parties. Copyright marking is seen as a partial solution [7, 9] to these problems. The mark can be embedded in any legal versions and will therefore be present in any copies made. This helps the copyright owner [1, 6, 8] to identify who has an illegal [10, 11] copy.

In general, there is a tradeoff between the watermarks embedding [11] strength (the watermark robustness) and quality (the watermark invisibility). Increased robustness

requires a stronger embedding, which in turn increases the visual degradation of the images. The proposed watermarking scheme adopts a color image as the watermark so human eyes can easily verify the extraction of this visually meaningful watermark. In general, a color image can provide more perceptual information i.e. sufficient evidence against any illegal copyright invasion.

This paper is organized as follows: Two-Dimensional Discrete Z-Transform has been presented with expression evaluated and simplified. The insertion and extraction technique for embedding the authenticating image in the carrier image has been introduced with suitable algorithm and example. The result of the proposed technique SSCIAZ compared with the existing Reversible data hiding based on block median preservation (RDHBBMP [13]) watermarking method in terms of visual interpretation, MSE, PSNR in dB and IF.

The techniques used in this paper includes two dimensional discrete Z-Transform and two dimensional discrete inverse Z-Transform represented as

1.1 Two Dimensional Z Transform

A function $f(n1, n2)$ can be represented in Z-Transform as

$$f(z1, z2) = \sum_{n1=-\infty}^{\infty} \sum_{n2=-\infty}^{\infty} f(n1, n2)z1^{-n1}z2^{-n2} \tag{1}$$

where $z1$ and $z2$ are both complex numbers consisting of real and an imaginary parts. Since $z1$ and $z2$ are complex numbers, Let $z1=e^{j\omega1\pi}$ and $z2=e^{j\omega2\pi}$, Where $e^{j\theta} = \cos\theta + j\sin\theta$. Substituting the values of $z1$ and $z2$ in equation (1), the equation becomes the discrete form of Two Dimensional Z Transformation equation.

$$f(e^{j\omega1\pi}, e^{j\omega2\pi}) = \sum_{n1=-\infty}^{\infty} \sum_{n2=-\infty}^{\infty} f(n1, n2)e^{j\omega1\pi^{-n1}}e^{j\omega2\pi^{-n2}}$$

$$\text{Or } f(\omega1, \omega2) = \sum_{n1=-\infty}^{\infty} \sum_{n2=-\infty}^{\infty} f(n1, n2)e^{-j\pi(n1\omega1+n2\omega2)} \tag{2}$$

1.2 Two Dimensional Inverse Z Transform

The Inverse Z-Transform of a function $f(n1, n2)$ is represented as

$$f(n1, n2) = \left(\frac{1}{2\pi j}\right)^2 \iint f(z1, z2)z1^{n1-1}z2^{n2-1} dz1dz2 \tag{3}$$

Where $f(n1, n2)$ be a function and $f(z1, z2)$ be the Z-Transform of the function $f(n1, n2)$.

1.3 Derivation of Inverse Z Transform from Continuous to Discrete Form

Since z_1 and z_2 are complex numbers, Let $z_1=e^{j\omega_1\pi}$ and $z_2=e^{j\omega_2\pi}$, where $e^{j\omega\theta} = \cos\omega\theta + j\sin\omega\theta$. Substituting the values of z_1 and z_2 in equation (3), we have a discrete form of inverse Z Transform for two dimensions. Now $z_1=e^{j\omega_1\pi}$, differentiating this with respect to ω_1 we get $\frac{dz_1}{d\omega_1} = e^{j\omega_1\pi}j\pi$, therefore $dz_1=e^{j\omega_1\pi}j\pi d\omega_1$ and $z_2=e^{j\omega_2\pi}$, differentiating this with respect to ω_2 we get $\frac{dz_2}{d\omega_2} = e^{j\omega_2\pi}j\pi$, therefore $dz_2=e^{j\omega_2\pi}j\pi d\omega_2$. The equation (3) becomes from the above derivation is

$$f(n_1, n_2) = \left(\frac{1}{2\pi j}\right)^2 \iint f(e^{j\omega_1\pi}, e^{j\omega_2\pi})e^{j\omega_1\pi n_1-1} e^{j\omega_2\pi n_2-1} e^{j\omega_1\pi}j\pi d\omega_1 e^{j\omega_2\pi}j\pi d\omega_2$$

The discrete form of this equation is as follows

$$f(n_1, n_2) = \frac{1}{4} \sum_{\omega_1=-1}^1 \sum_{\omega_2=-1}^1 f(\omega_1, \omega_2)e^{j\pi(n_1\omega_1+n_2\omega_2)} \tag{4}$$

The equation (4) is the discrete form of Two Dimensional Inverse Z Transform.

2 The Technique

The Insertion of the authenticating image is performed in the Z-Domain i.e. the domain obtained after performing the Z-Transform on 2 x 2 sub-image matrix of the original image matrix one by one. Hence, in order to perform the insertion operation of the authenticating image into converted original image byte. Bits from authenticating image are embedded in single bit position under each byte of the source image. The authenticating image pixels are read and are converted into binary values and each binary bit is inserted into one pixel of the original image into which the watermark is supposed to be embedded. Point of insertion of a bit is obtained by computing the Fibonacci series and then taking the LSB two bits as the position of insertion of the bit to be embedded in the image.

Let C_w be the pixel value in Watermarked color image and C be the original pixel value of the digital image to be embedded. Let $b[i]$ be the bit to be embedded in pixel C . The embedding or coding step and the detection scheme or the decoding step is as follows:

2.1 Insertion Algorithm

Input: A carrier image and authenticating message/image.

Output: An authenticated image.

Method: Embedding has been performed on the integer values only while the floating point part has been made intact and has been added after embedding the watermarking bits in the integer part of the pixels values of the source image.

1. Read Image type, dimensions and maximum intensity from source image and write in the output image.
2. Repeat until all pixels have been read from the source image file.
 - 2.1 Repeatedly Take 2×2 blocks of pixels from the matrix at the left and perform Z-Transform of the block of pixels until all pixels in the matrix have been taken.
 - 2.2 Compute the Fibonacci series and generate the positions using the two LSB bits of the generated number where the watermark bits will be embedded. The Fibonacci series will be repeated after taking a specific number of terms.
 - 2.3 Read the authenticating image i.e. watermark.
 - 2.4 Embed the watermark bits in the source image in the position specified by the number generated from the Fibonacci series.
 - 2.5 Compute the Inverse Z-Transform of the 2×2 block of pixels.
 - 2.6 If any pixel is found to be of negative value, the maximum negative number is stored and added in the watermarked pixel values such that there is no effect on the bit position where the watermark bit is embedded.
 - 2.7 Compute the Inverse Z-Transform of the block of pixels and the numbers obtained is guaranteed to be of positive values.
 - 2.8 Repeat the steps from 2.1 to 2.7 until all pixels have been transformed.
3. Stop.

2.2 Extraction Algorithm

The Extraction of the authenticating image is performed in the Z-Domain i.e. the domain obtained after performing the Z-Transform of the embedded image. Hence, in order to perform the extraction operation of the authenticating image from the embedded image, the embedded image is first converted using Z-Transform.

A masking based detection scheme has been proposed to retrieve the embedded watermark from a color carrier image. Bits from authenticating image have been embedded in single bit position under each byte of the source image by choosing a standard 2×2 mask in row major order. In case of retrieval of the authenticating image, we will have only one extracted bit from each pixel of the embedded image. Point of extraction of a bit is obtained by computing the Fibonacci series and then taking the LSB two bits as the position of extraction of the bit from the embedded image.

Input: Authenticated image.

Output: The original image, authenticating message/image.

Method: Extraction has been performed on the integer values only while the floating point part has been made intact and has been added after extracting the security bits from the integer part of the pixels values of the source image.

1. Read Image type, dimensions and maximum intensity from embedded image and skip writing in the output image.
2. Repeat until all pixels have been read from the embedded image.
 - 2.1. Repeatedly Take 2×2 blocks of pixels from the matrix at the left and perform Z-Transform of the block of pixels until all pixels in the matrix have been taken.

- 2.2. Compute the Fibonacci series and generate the positions using the two LSB bits of the generated number where the watermark bits have been embedded. The Fibonacci series will be repeated after taking a specific number of terms so as to avoid core dumb and overflow conditions.
 - 2.3. Calculate the embedded bits from the embedded image from the position specified by the number generated from the Fibonacci series.
 - 2.4. Convert each 8 bits of 0's and 1's into decimal value and write the value in the output image.
 - 2.5. Repeat the steps from 2.1 to 2.6 until all pixels have been transformed and embedded bits have been calculated.
3. Stop.

3 Result Comparison and Analysis

This section represent the results, discussion and a comparative study of the proposed technique SSCIAZ with the DCT-based watermarking method, QFT-based and Spatio Chromatic DFT-based watermarking method in terms of visual interpretation, image fidelity (IF), and peak signal-to-noise ratio (PSNR) analysis and mean square error (MSE). In order to test the robustness of the scheme SSCIAZ, the technique is applied on more than 50 PPM colour images from which it may be revealed that the algorithm may overcome any type of attack like visual attack and statistical attack. Experimental set up for preparing result is any type of PC with 2.00 GHz and above processing speed, 2 GB primary memory and Fedora 6 or above version of OS with gimp application. Distinguishing of carrier and embedded image from human visual system is quite difficult. In this section some statistical and mathematical analysis is given. The original carrier images 'Airplane', 'Baboon', 'Lenna' and 'Oakland' are shown in fig 1a, 1b, 1c and 1d. The dimension of each carrier colour images is 512 x 512 and the dimension of the authenticating colour image (Fig. 1m) is shown in table 1 and 2. Embedded colour images are shown in Fig 1e, 1f, 1g and 1h using SSCIAZ. Single bit of secrete data is embedded in each carrier image byte. The magnified versions of different images have been shown on Fig. 1i, 1j, 1k and 1l.

Peak signal-to-noise ratio (PSNR) is used to evaluate qualities of the stego-images. Table 1 and 2 shows two levels of authenticating data byte embedding which is defined by EL=0 and EL=1 based on PSNR values. Table 2 shows the PSNR values for comparative studies of SSCIAZ and Reversible data hiding based on block median preservation (RDHBBMP) and also the enhancement in terms of hiding capacity of secrete data and PSNR in dB. The average enhancement of secrete data embedding is 34553 bits in SSCIAZ than the existing technique RDHBBMP and also .14 dB of PSNR in EL=0. But in EL=1 the average enhancement of secrete data embedding is 137697.5 bits in SSCIAZ than the existing technique RDHBBMP and also .33 dB of PSNR. In all the existing technique the PSNRs are low, means bit-error rate are high but in the proposed scheme more bytes of authenticating data can be embedded and the PSNR values are significantly high, means bit-error rate is low. The average improvement is shown in Table 2. Table 3 shows the better PSNR values than other exiting techniques like DCT-based [10] watermarking, QFT-based [11] watermarking, and SCDFT-based [12] watermarking in frequency domain. Capacities

of existing techniques are 3840 bytes and the PSNR values are 30.1024 dB, 30.9283 dB, and 30.4046 dB in SCDFT, QFT, and DCT respectively. Whereas the capacity of SSCIAZ is 8112 bytes and PSNR is 49.89 dB and which is fully recoverable. 4272 bytes more secrete data embedding is possible in SSCIAZ technique than existing techniques with average 19 dB more PSNR values.

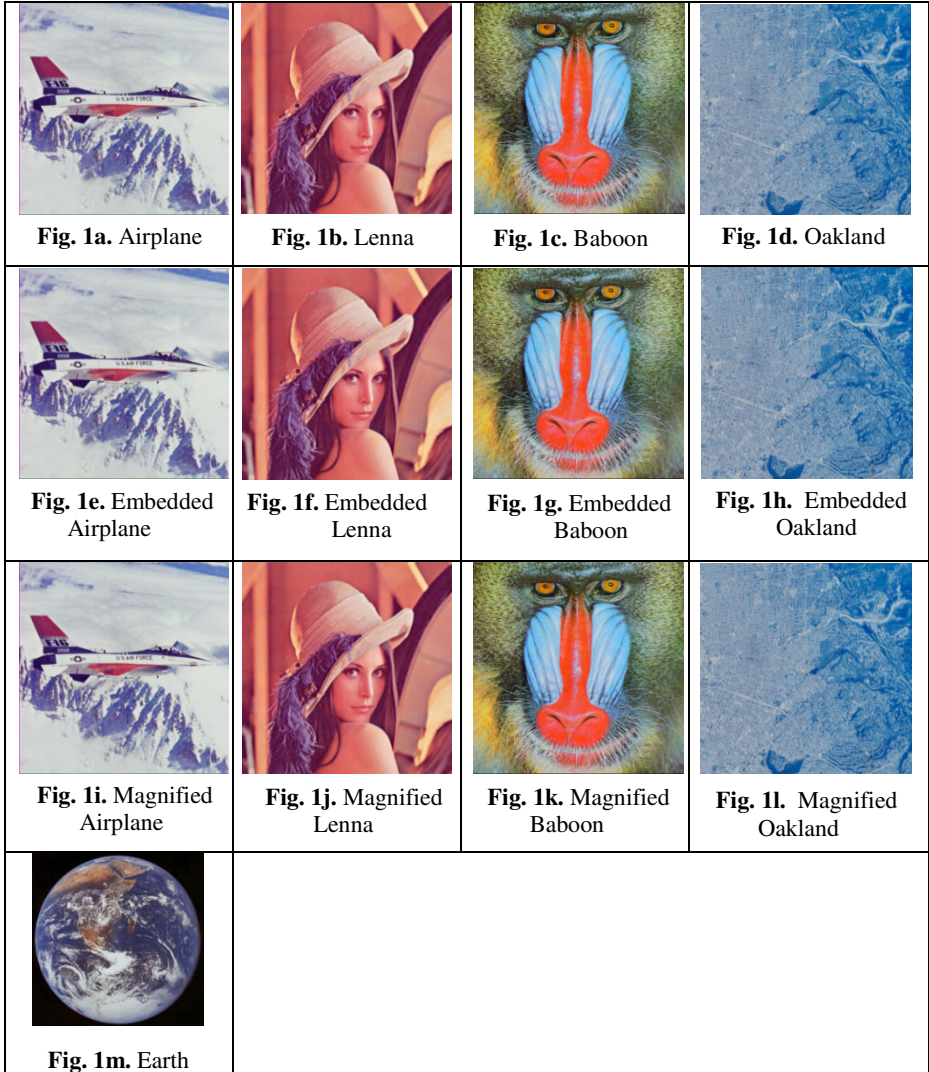


Fig. 1. Visual interpretation of embedded image using SSCIZ and corresponding magnified images after embedding

Table 1. Capacities and PSNR values of SSCIAZ

Test images	Indicator	EL=0	EL=1
Baboon	C(bits)	60,000	194400
	PSNR	50.16	45.18
Oakland	C(bits)	60,000	194400
	PSNR	50.17	45.19
Peppers	C(bits)	60,000	194400
	PSNR	50.21	45.22
Average Image	C(bits)	60,000	194400
	PSNR	50.18	45.20

Table 2. Results and comparison in capacities and PSNR of SSCIAZ and RDHBBMP

Test images	Indicator	EL=0		EL=1	
		RDHBBMP	SSCIAZ	RDHBBMP	SSCIAZ
Lena	C(bits)	26,465	64,896	71,769	2,16,600
	PSNR	49.68	49.89	44.35	44.76
Airplane	C(bits)	36,221	64,896	86,036	2,16,600
	PSNR	49.80	49.87	44.64	44.89
Average Image	ΔCa	34553		137697.5	
	$\Delta PSNRa$	0.14		0.33	

Table 3. Results and comparison in capacities and PSNR of SSCIAZ and DCT, QFT, SCDFT [12]

Technique	Capacity (bytes)	PSNR in dB
SCDFT	3840	30.1024
QFT	3840	30.9283
DCT	3840	30.4046
SSCIAZ	8112	49.8900

4 Conclusion

SSCIAZ technique is an image authentication process to enhance the security compared to the existing algorithms. Authentication is done by embedding secret data embedding in each mask of carrier image byte is possible. In compare to Reversible data hiding based on block median preservation, SSCIAZ algorithm is applicable for any types of colour image authentication and strength is high. The PSNR is high and more bytes of authenticating bits can be embedded in the carrier images. The watermark embedded in this method is very hard to detect due to unknown insertion position of the authenticating bits in the carrier image. So, the proposed technique SSCIAZ also provides security from all possible attacks.

Acknowledgement. The author expresses the deep sense of gratitude to the Dept. of Engineering and Technological Studies, University of Kalyani, West Bengal, India, where the work has been carried out.

References

1. Radhakrishnan, R., Kharrazi, M., Menon, N.: Data Masking: A new approach for steganography. *Journal of VLSI Signal Processing* 41, 293–303 (2005)
2. EL-Emam, N.N.: Hiding a large Amount of data with High Security Using Steganography Algorithm. *Journal of Computer Science* 3(4), 223–232 (2007) ISSN 1549-3636
3. Amin, P., Lue, N., Subbalakshmi, K.: Statistically secure digital image data hiding. In: *IEEE Multimedia Signal Processing MMSP 2005*, Shanghai, China, pp. 1–4 (October 2005)
4. Pavan, S., Gangadharpalli, S., Sridhar, V.: Multivariate entropy detector based hybrid image registration algorithm. In: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Philadelphia, Pennsylvania, USA, pp. 18–23 (March 2005)
5. Al-Hamami, A.H., Al-Ani, S.A.: A New Approach for Authentication Technique. *Journal of Computer Science* 1(1), 103–106 (2005) ISSN 1549-3636
6. Ker, A.: Steganalysis of Embedding in Two Least-Significant Bits. *IEEE Transaction on Information Forensics and Security* 2(1), 46–54 (2008) ISSN 1556-6013
7. Yang, C., Liu, F., Luo, X., Liu, B.: Steganalysis Frameworks of Embedding in Multiple Least Significant Bits. *IEEE Transaction on Information Forensics and Security* 3(4), 662–672 (2008) ISSN 1556-6013
8. Wu, H.C., Wu, N.I., Tsai, C.S., Hwang, M.S.: Image steganographic scheme based on pixel-value differencing and LSB replacement methods. *Proc. Inst. Elect. Eng., Vis. Images Signal Processing* 152(5), 611–615 (2005)
9. Yang, C.H., Weng, C.Y., Wang, S.J., Sun, H.M.: Adaptive Data Hiding in edge areas of Images With Spatial LSB Domain Systems. *IEEE Transaction on Information Forensics and Security* 3(3), 488–497 (2008) ISSN 1556-6013
10. Ahmidi, N., Safabakhsh, R.: A novel DCT-based approach for secure color image watermarking. In: *Proc. Int. Conf. Information Technology: Coding and Computing*, vol. 2, pp. 709–713 (2004)
11. Bas, P., Biham, N.L., Chassery, J.: Color watermarking using quaternion Fourier transformation. In: *Proc. ICASSP*, Hong Kong, China, pp. 521–524 (June 2003)
12. Tsui, T.T., Zhang, X.-P., Androutsos, D.: Color Image Watermarking Using Multidimensional Fourier Transformation. *IEEE Trans. on Info. Forensics and Security* 3(1), 16–28 (2008)
13. Luo, H., Yu, F.-X., Chen, H., Huang, Z.-L., Li, H., Wang, P.-H.: Reversible data hiding based on block median preservation. *Information Sciences* 181, 308–328 (2011)

Heart Disease Diagnosis Using Machine Learning Algorithm

Shashikant U. Ghumbre¹ and Ashok A. Ghatol²

¹ Computer Engineering Department,
College of Engineering Pune, Pune, Maharashtra, India
shashi.ghumbre@gmail.com

² Dr. B.A.T.University, Lonere, Maharashtra, India
vc_2005@rediffmail.com

Abstract. Recent advances in computing and developments in technology have facilitated the routine collection and storage of medical data that can be used to support medical decisions. However, in most countries, there is a first need for collecting and organizing patient's data in digitized form. Then, the collected data are to be analyzed in order for a medical decision to be drawn, whether this involves diagnosis, prediction, course of treatment, or signal and image analysis. In this paper, India centric dataset is used for Heart disease diagnosis. The correct diagnosis performance of the automatic diagnosis system is estimated by using classification accuracy, sensitivity and specificity analysis. The study shows that, the SVM with Sequential Minimization Optimization learning algorithm have better choice for medical disease diagnosis application.

Keywords: Decision Support System, Support Vector Machine, Heart Disease, Machine Learning.

1 Introduction

The application of machine learning methods in medical field is the subject of considerable ongoing research, which mainly concentrates on modeling some of the human actions or thinking processes and recognizing diseases from a variety of input sources. Other application areas are knowledge discovery [10] and biomedical systems, which include genetics and DNA analysis [1, 12]. The design may also be influenced by the desired performance on one or more specific classes of the problem instead of the overall performance. This is usual in most medical tasks as a different degree of significance may be required for the system's performance on each class. The computer programs or machine learning techniques can be used to reduce the mortality rate, improve the accuracy in disease diagnosis and mainly reduce the diagnosis time. The advancement in computer technology and communication encourages health-care providers to work using the Internet or Telemedicine technology [9,13,14].

In a medical diagnosis problem, what is needed is a set of examples or attributes that are representative of all the variations of the disease. The examples need to be selected very carefully if the system is to perform reliably and efficiently. The fact that there is no need to provide a specific algorithm on how to identify the disease, presents a major advantage over the application of machine learning methods to this

type of problems. However, development of artificial intelligence systems for medical decision making problems is not a trivial task. Difficulties include the acquisition, collection and organization of the data that will be used for training the system. This becomes a major problem especially when the system requires large data sets over long periods of time, which in most cases are not available due to the lack of an efficient recording system. The above mentioned problems or the existing procedures involved in the medical task may not be the only factors affecting the design of a Decision Support System (DSS). The design may also be influenced by the desired performance on one or more specific classes of the problem instead of the overall performance. This is usual in most medical tasks as a different degree of significance may be required for the system's performance on each class. For example, in a heart disease diagnosis task, it is necessary for the accuracy on healthy patients to be as high as possible, as a misclassification in this category may result in a healthy patient going under treatment for no reason. The balance of the system's performance between different classes could vary and is largely dependent on the medical problem itself and the collected data. In addition, in most of the countries, insufficient numbers of medical specialist have increased the mortality of patients suffering from various diseases. Heart diseases have emerged as the number one killer in both urban and rural areas in most of the countries. As of 2010, it is the leading cause of death in the U.S., England and Canada, accounting for 25.4% of the total deaths in the United States. Similar situation is found rest of the countries all over the world. In case of heart disease time is very crucial to get correct diagnosis in early stage[37]. It is observed that, in many cases due to wrong diagnosis or trial/error procedure for diagnosis leads to patient health compromise. The dearth of medical specialists and/or wrong diagnosis procedure will never be overcome within a short period of time [3,4]. Patient having chest pain complaint may undergo unnecessary treatment or admitted in the hospital. In most of the developing countries specialists are not widely available for the diagnosis. Hence, such automated system can help to medical community to assist doctor for the accurate diagnosis well in advance.

The rest of the paper is organized as follows: Section 2 briefly reviews some prior works on machine learning techniques in medical Diagnosis. Section 3 briefly describes the heart disease diagnosis and the proposed Decision Support System (DSS) and its techniques used are discussed. Section 4 details the use of Support Vector Machine in medicine. The experimental results are given in Section 5. Section 6 concludes the paper.

2 Literature Review

Zhi-Hua Zhou and Yuan Jiang [4] have proposed an approach named C4.5 Rule-PANE, which gracefully combines the advantages of artificial neural network ensemble and rule induction. A specific rule induction approach, i.e. C4.5 Rule, is used to learn rules from the new training data set. Case studies on diabetes, hepatitis, and breast cancer show that C4.5 Rule-PANE could generate rules with strong generalization ability, which profits from artificial neural network ensemble, and strong comprehensibility, which profits from rule induction.

Leung et al., [32] have presented a data mining framework for biological data sets. And it has been applied to the Hepatitis B Virus DNA data sets which are real world data. Their method has good performance using the fuzzy measure and the nonlinear

integral, since the non additivity of the fuzzy measure reflects the importance of the feature attributes, as well as their inherent interactions.

Saangyong Uhm et al., [34] have presented the machine learning techniques, SVM, decision tree, and decision rule to predict the susceptibility of the liver disease, chronic hepatitis from single nucleotide polymorphism (SNP) data. The experimental results have shown that decision rule is able to distinguish chronic hepatitis from normal with the maximum accuracy of 73.20%, whereas SVM is with 67.53% and decision tree is with 72.68%. It is also shown that decision tree and decision rule is potential tools to predict the susceptibility to chronic hepatitis from SNP data.

Ozyilmaz, L. and Yildirim, T [35] have presented three neural network algorithms for diagnosis of hepatitis diseases. The results were compared with some statistical methods used in a previous work. Their results have shown that using a hybrid network CSFNN that combines MLP and RBF is more reliable for the diagnosis. Thus the compared results have shown that neural networks can be used in the problem of diagnosis for hepatitis diseases as efficiently as statistical methods.

3 Heart Disease Diagnosis

3.1 Heart Disease

The heart, shown in Figure 1, is actually two separate pumps: a *right heart* that pumps blood through the lungs, and a *left heart* that pumps blood through the peripheral organs. In turn, each of these hearts is a pulsatile two-chamber pump composed of an *atrium* and a *ventricle*. Each atrium is a weak primer pump for the ventricle, helping to move blood into the ventricle. The ventricles then supply the main pumping force that propels the blood either through the pulmonary circulation by the right ventricle or through the peripheral circulation by the left ventricle.

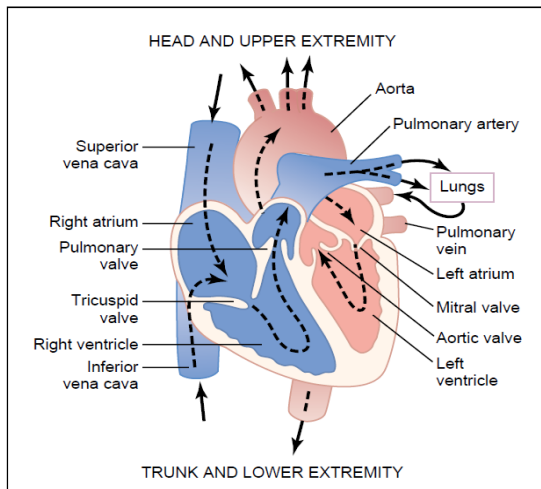


Fig. 1. Structure of the heart, and course of blood flow through the heart chambers and heart valves

The cardiac events that occur from the beginning of one heartbeat to the beginning of the next are called the *cardiac cycle*. The cardiac cycle consists of a period of relaxation called *diastole*, during which the heart fills with blood, followed by a period of contraction called *systole*. Blood normally flows continually from the great veins into the atria; about 80 per cent of the blood flows directly through the atria into the ventricles even before the atria contract. Then, atrial contraction usually causes an additional 20 per cent filling of the ventricles. Therefore, the atria simply function as primer pumps that increase the ventricular pumping effectiveness as much as 20 per cent. When the atria fail to function, the difference is unlikely to be noticed unless a person exercises; then acute signs of heart failure occasionally develop, especially shortness of breath. The *P, Q, R, S, and T waves* in electrocardiogram are electrical voltages generated by the heart and recorded by the electrocardiograph from the surface of the body [36]. The choice of which tests to perform depends on several things. These include a patient's risk factors, history of heart problems, current symptoms and the doctor's interpretation of these factors. The cardiovascular problems may experienced with ordinary physical activity causes symptoms: Undue fatigue, Palpitations-the sensation that heart is skipping a beat or beating too rapidly, Dyspnea-difficult or labored breathing, Anginal pain -chest discomfort from increased activity. The doctor diagnose the heart attack with, review the patient's complete medical history, physical examination, use an electrocardiogram (ECG) to discover any abnormalities caused by damage to the heart, sometimes use a blood test to detect abnormal levels of certain enzymes in the bloodstream. Blood tests confirm (or refute) suspicions raised in the early stages of evaluation that may occur in an emergency room, intensive care unit or urgent care setting. These tests are sometimes called heart damage markers or cardiac enzymes [37].

3.2 Automated Heart Disease Diagnosis

The first stage of this diagnosis study decodes the real and typical doctor-based diagnosis procedure for the particular Heart disease. During data collection, number of visits to the hospitals and discussions with medical practitioners has been carried out. Table: 1 shows the Heart disease symptoms and tests which are usually observed during the diagnosis.

Table 1. Heart Disease Attributes

Symptoms and signs	Test results
Chest pain Types(Left and Right), Arm pain, backache, Sweating, Breathlessness, addiction, Diabetic, MAP, Pulse rate	ECG: ST Elevation, ST Depression, T Elevation, T Depression, Q waves, BSL, CK-MB

Based on medical records 214 instances in total are selected for the automated diagnosis. Dataset consists of 19 different attributes with four class distribution: 0- Myalgia, 1- Myocardial Infarction (MI), 2- Ischemic Heart Disease (IHD), 3- Unstable Angina (UA). For this study binary classification problem is considered for

Heart disease diagnosis, in which 78 instances are belongs to 0 i.e. Myalgia (Normal), and 139 are considered as 1 i.e. Patient having Heart disease. In order to provide physicians with both structured questions and structured responses within medical domains of specialized knowledge or experience [17,25,26] medical expert systems have been developed. The advice of one or more medical experts, who also suggest the optimal questions to be considered, and provide the most accurate conclusions to be drawn from the answers the physician chooses, is used to embody the structure in the program. A trained Support Vector Machine (SVM), Multilayer Perceptron (MLP), Radial Basis Function Neural Network (RBF) and other methods are used to assume the evolution of the biological indicators. Once the patients' personal data is presented along with the results of the tests taken at the onset of the treatment and the postulated code of reaction, the evolution in time of the illness can be specified by the expert system.

4 Support Vector Machine

Support Vector Machine (SVM) is a category of universal feed forward networks like Radial-basis function networks, pioneered by Vapnik. SVM can be used for pattern classification and nonlinear regression. More precisely, the support vector machine is an approximate implementation of the method of structural risk minimization. This principle is based on the fact the error rate of a learning machine on test data is bounded by the sum of the training-error rate and term that depends on the Vapnik-Chervonenkis (VC) dimension [33]. The support vector machine can provide good generalization performance on pattern classification problem.

4.1 Optimal Hyperplane for Patterns

Consider the training sample $\{(x_i, y_i)\}_{i=1}^N$ where x_i is the input pattern for the i th instance and y_i is the corresponding target output. With pattern represented by the subset $y_i = +1$ and the pattern represented by the subset $y_i = -1$ are linearly separable. The equation in the form of a hyperplane that does the separation is

$$w^T x + b = 0 \quad (1)$$

where, x is an input vector, w is an adjustable weight vector, and b is a bias. Thus,

$$w^T x_i + b \geq 0 \quad \text{for } y_i = +1 \quad (2)$$

$$w^T x_i + b < 0 \quad \text{for } y_i = -1 \quad (3)$$

for a given weight vector w and a bias b , the separation between the hyperplane defined in eq. 1 and closest data point is called the margin of separation, denoted by ρ as shown in figure 2, the geometric construction of an optimal hyperplane for a two-dimensional input space.

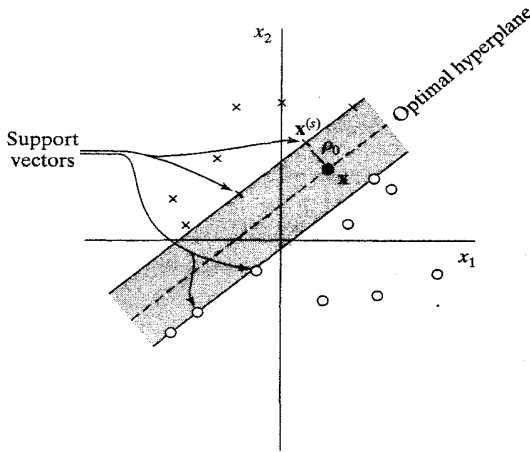


Fig. 2. Optimal Hyperplane for a two dimensional input space

The discriminant function gives an algebraic measure of the distance from x to the optimal hyperplane for the optimum values of the weight vector and bias, respectively.

$$g(x) = w_o^T x + b_o \tag{4}$$

5 Experimentations and Results

The diagnosis of the disease for a new patient to be performed on basis of dataset is facilitated by the first stage of data preparation and feature extraction. Initially, network size, sample size, model selection, and feature extraction are considered as key parameters for the study of a practical network design issues related to learning and generalization. During experimentation, it is observed that for such type of nonlinear classification problem unsupervised learning procedure for feature extraction is not appropriate because of nonlinear correlation structure. It is also observed that, correct and complete data collection procedure is the proper route for the selection of best classifier. Heart disease dataset is used for diagnosis with 214 instances of 3 types of predicted heart diseases. For binary classification problem with limited data size it is necessary to validate networks model with cross validation, hence 5-fold and 10-fold cross validation techniques are used on heart disease dataset. Table 2 and 3 shows comparative results of SVM, MLP, RBF, BayesNet, J48 and Rule, in which SVM gives promising results using 5-fold and 10-fold cross validation.

Table 2. Generalization performance using 5-fold cross validation

Classifier	Accuracy	Sensitivity	Specificity
SVM	85.51%	84.60%	88.50%
MLP	82.71%	85.30%	78.20%
RBF	82.24%	82.40%	82.10%
BayesNet	79.90%	77.20%	84.60%
J46	71.43%	73.54%	70.10%
Rule	68.90%	72.81%	69.94%

Table 3. Generalization performance using 10-fold cross validation

Classifier	Accuracy	Sensitivity	Specificity
SVM	85.05%	84.60%	85.90%
MLP	84.11%	87.50%	78.20%
RBF	82.71%	83.10%	82.10%
BayesNet	80.37%	77.20%	85.90%
J48	76.65%	73.80%	74.10%
Rule	71.16%	67.90%	72.80%

6 Conclusions

Medical diagnosis has become highly attributed with the development of technology lately. Furthermore the computer and communication tools have improved the medical practice implementation to a greater extent. Though the Artificial neural network ensemble is a powerful learning technique that could aid in the remarkable improvement in the generalization ability of neural learning systems its lucidity is worse than that of a single artificial neural network thereby deterring its wide recognition among the medical practitioners. In this paper we have projected a Decision Support System for the diagnosis of Heart disease by means of radial basis function network structure and Support Vector Machine. Therefore the diagnosis of Heart disease is carried out utilizing different data samples from diverse patients and the results have denoted that SVM with Sequential minimize optimization is equivalently good as the ANN and other models in the diagnosis of Heart disease. The classification accuracy, sensitivity, and specificity of the SVM has been found to be high thus making it a good option for the diagnosis.

References

1. Brause, R.W.: Medical Analysis and Diagnosis by Neural Networks. In: Proceedings of Medical Data Analysis, October 8-9, vol. 20, pp. 1–13. Springer, Heidelberg (2001)
2. Gerard Wolff, J.: Medical diagnosis as pattern recognition in a framework of information compression by multiple alignment, unification and search. Decision Support Systems 42(2), 608–625 (2006)

3. Steimann, F., Adlassnig, K.P.: Fuzzy medical diagnosis. In: Rupini, E., Bonissone, P., Pedrycz, W. (eds.) *Handbook of Fuzzy Computation*. Oxford University Press and Institute of Physics Publishing, Bristol (1998)
4. Zhou, Z.-H., Jiang, Y.: Medical Diagnosis with C4.5 Rule Proceeded by Artificial Neural Network Ensemble. *IEEE Transactions on Information Technology in Biomedicine* 7(1), 37–42 (2003)
5. Richards, G., Rayward-Smith, V.J., Sönksen, P.H., Carey, S., Weng, C.: Data mining for indicators of early mortality in a database of clinical records. *Artificial Intelligence in Medicine* 22(3), 215–231 (2000)
6. Ćosić, D., Lončarić, S.: Rule-Based Labeling of CT Head Image. In: *Proceedings of the 6th European Conf. of AI in Medicine Europe AIME 1997*, pp. 453–456 (1997)
7. Duch, W., Adamczak, R., Grąbczewski, K., Żal, G., Hayashi, Y.: Fuzzy and crisp logical rule extraction methods in application to medical data. In: Szczepaniak, P.S., Lisboa, P.J.G., Kacprzyk, J. (eds.) *Fuzzy systems in Medicine*, pp. 593–616. Physica - Verlag, Springer, Heidelberg (1999)
8. Stevens, A., Lowe, J.S., Young, B.: *Wheater's Basic Histopathology: a color atlas and text*, 4th edn., p. 315. Churchill Livingstone (2003) ISBN-10 0-4430-7001-6
9. Manickam, S., Abidi, S.S.R.: Experienced Based Medical Diagnostics System Over The World Wide Web (WWW). In: *Proceedings of The First National Conference on Artificial Intelligence Application In Industry*, Kuala Lumpur, pp. 47–56 (1999)
10. Alexopoulos, E., Dounias, G.D., Vemmos, K.: Medical Diagnosis of Stroke Using Inductive Machine Learning. In: *Machine Learning and Applications: Machine Learning in Medical Applications*, Chania, Greece, pp. 20–23 (1999)
11. Shortliffe, E.H.: *Computer Programs to Support Clinical Decision Making. Readings in Uncertain Reasoning*, 161–166 (1990) ISBN:1-55860-125-2
12. Neves, J., Alves, V., Nelas, L., Romeu, A., Basto, S.: An Information System, That Supports Knowledge Discovery and Data Mining in Medical Imaging. In: *Machine Learning and Applications: Machine Learning in Medical Applications*, Chania, Greece, pp. 37–42 (1999)
13. Cunningham, P., Carney, J., Jacob, S.: Stability problems with artificial neural networks and the ensemble solution. *Artificial Intelligence in Medicine* 20(3), 217–225 (2000)
14. Kononenko: Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine* 23(1), 89–109 (2001)
15. Zhou, Z.-H., Jiang, Y., Yang, Y.B., Chen, S.F.: Lung cancer cell identification based on artificial neural network ensembles. *Artificial Intelligence in Medicine* 24(1), 25–36 (2002)
16. Pranckeviciene, E.: Finding Similarities Between An Activity of the Different EEG's by means of a Single layer Perceptron. In: *Machine Learning and Applications: Machine Learning in Medical Applications*, Chania, Greece, pp. 49–52 (1999)
17. Luger, G.F., Stubblefield, W.A.: *Artificial intelligence and the design of expert systems*. Benjamin/Cummings Publ. Co., Redwood City (1989)
18. Dimitrios Siganos: *Neural Networks in Medicine* (1995), from http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol2/ds12/article2.html
19. Wolff, J.G.: Medical Diagnosis as Pattern Recognition in a Framework of Information Compression by Multiple Alignment, Unification and Search. *Decision Support Systems* 42(2), 608–625 (2006)
20. Sasikala, K.R., Petrou, M., Kittler, J.: Fuzzy classification with a GIS as an aid to decision making. *EARSel Advances in Remote Sensing* 4, 97–105 (1996)
21. Tou, J.T., Gonzalez, R.C.: *Pattern Recognition Principles*. Addison – Wesley, Reading (1974)
22. Moody, J., Darken, C.J.: Fast learning in networks of locally tuned processing units. *Neural Computation* 2, 281–294 (1989)

23. Hanson, S.J., Burr, D.J.: Minkowski-r back propagation: learning in connectionist models with non-Euclidean error signals. In: *Neural Information Processing Systems*, pp. 348–357. American Institute of Physics, New York (1988)
24. Poggio, T., Girosi, F.: Regularization algorithms for learning that are equivalent to multilayer networks. *Science* 247, 978–982 (1990)
25. Poggio, T., Girosi, F.: Networks for approximation and learning. *Proceedings of the IEEE* 78(9), 1481–1497 (1990)
26. Hartman, E.J., Keeler, J.D., Kowalski, J.M.: Layered neural networks with Gaussian hidden units as universal approximators. *Neural Computation* 2, 210–215 (1990)
27. Park, J., Sandberg, I.W.: Universal approximation using radial basis function networks. *Neural Computation* 3, 246–257 (1991)
28. Park, J., Sandberg, I.W.: Approximation and radial basis function networks. *Neural Computation* 5, 305–316 (1993)
29. Venkatesan, P., Anitha, S.: Application of a radial basis function neural network for diagnosis of diabetes mellitus. *Current Science* 91(9), 1195–1199 (2006)
30. Lin, T.-C., Kuo, M.-J., Chen, Y.-C.: Frequency Domain Analog Circuit Fault Diagnosis Based on Radial Basis Function Neural Network. In: *International Conference on Communications, Circuits and Systems, ICCAS 2004, June 27-29, vol. 2*, pp. 1183–1185 (2004)
31. Bhatia, S., Prakash, P., Pillai, G.N.: SVM based Decision Support System for Heart Disease Classification with Integer-coded Genetic Algorithm to select critical features. In: *Proceedings of the World Congress on Engineering and Computer Science, San Francisco, USA*, pp. 34–38 (2008)
32. Leung, K.S., Ng, Y.T., Lee, K.H., Chan, L.Y., Tsui, K.W., Mok, T., Tse, C.H., Sung, J.: Data Mining on DNA Sequences of Hepatitis B Virus by Nonlinear Integrals. In: *Proceedings Taiwan-Japan Symposium on Fuzzy Systems & Innovational Computing, 3rd Meeting, Japan*, pp. 1–10 (August 2006)
33. Haykin, S.: *Neural Networks*, 2nd edn. Prentice-Hall (2003)
34. Uhm, S., Kim, D.-H., Kim, J., Cho, S.W., Cheong, J.Y.: Chronic Hepatitis Classification Using SNP Data and Data Mining Techniques. In: *Frontiers in the Convergence of Bioscience and Information Technologies, FBIT 2007, October 11-13*, pp. 81–86 (2007)
35. Ozyilmaz, L., Yildirim, T.: Artificial neural networks for diagnosis of hepatitis disease. In: *Proceedings of the International Joint Conference on Neural Networks, July 20-24, vol. 1*, pp. 586–589
36. Gyton, C., Hall, J.E.: *Textbook of Medical Physiology*, 11th edn. Elsevier (2006)
37. <http://www.heart.org>

Position Determination and Face Detection Using Image Processing Techniques and SVM Classifier

Gogula Suvarna Kumar¹, P.V.G.D. Prasad Reddy²,
Sumit Gupta¹, and Ravva Anil Kumar¹

¹ Department of Computer Science Engineering
Maharaj Viajayam Gajapathi Raj College of Engineering
Vizianagaram, Andhra Pradesh, India

{gsk, prasadreddy.vizag, sumit108, ravvaanilkumar}@gmail.com

² Department of Computer Science and Systems Engineering
Andhra University Vishakapatnam, India

Abstract. In this paper, an improved algorithm for detecting the position of a person in controlled environments using the face detection algorithm is proposed. This algorithm ingeniously combines different face detection, occlusion detection algorithms and SVM classifier. A class room environment with thirty students is used along with some constraints such as position of the camera being fixed in a way that covers all the students, the static student's position and the class environment with the fixed lighting conditions. The students are treated as classes in this technique. For every class, a set of 6 attributes are derived and updated in a database. The image is given as an input to the face detection algorithm to detect some of the faces. Some faces are not detected because of occlusion, so an occlusion detection technique is implemented to detect all the occluded faces. In the training phase, a set of four images with the entire thirty students taken in four different days is used. Therefore a database of total 120 set of records with 6 attributes is used.

Keywords: Face Detection, Occlusion Detection, SVM, EMD.

1 Introduction

In the last decade, it can be observed that many algorithms were developed in image processing for face recognition [1] and face detection [2] but there was no algorithm for detecting a position in an image. The position detection in the controlled environment can be used in many environments where the positions are fixed. Some environments like seminar halls, conference halls, Auditoriums etc. And we can apply the same algorithm for some environments where the positions are reserved. With the help of this position detection algorithm many applications can be developed like automatic attendance system [3] in controlled environments. This technique can also be used for improving the face detection technique although no face detection technique will give the 100 % accuracy in its respective detection. Some of the faces in an image are blurred, not clear, occluded, the normal face detection algorithm will not detect those faces but this technique detects most of the faces in an image. First

the image is given as an input to the basic face detection algorithm followed by giving the acquired result of the above as an input to the occlusion detection [4] and then this technique is applied.

A large number of face detection algorithms are derived from neural network approach, algorithmic approach [5] and some image morphological techniques [6]. However most of the works concentrate on single face detection, with some constrained environments. In this proposed technique, a face detection algorithm by using local SMQT features and split up snow classifier [7] and an occlusion detection algorithm is used and from the result obtained the attributes are derived and updated in the database. A face detection algorithm is used which is implemented using the Local SMQT Features and split up Snow Classifier.

2 Existing System

Many face detection algorithms are derived. Some of the face detection algorithms use neural network approach, algorithmic approach and some image morphological techniques. With the help of these existing face detection algorithms, 80% of accuracy is obtained. When the result is given as an input to the occlusion detection algorithm, the accuracy is improved to 95% accuracy. Then the above acquired result is given as an input to the proposed technique. Face detection algorithm using Local SMQT Features and Split up Snow classifier is explained in the next paragraph.

Illumination and sensor variation are major concerns in visual object detection. It is desirable to transform the raw illumination and sensor varying image so the information only contains the structures of the object. Some techniques previously proposed to reduce this variation are Histogram Equalization (HE), variants of Local Binary Patterns (LBP) [8] and the Modified Census Transform (MCT) [9]. HE is a computationally expensive operation in comparison to LBP and MCT, however LBP and MCT are typically restricted to extract only binary patterns in a local area. The Successive Mean Quantization Transform (SMQT) [10] can be viewed as a tunable tradeoff between the number of quantization levels in the result and the computational load.

The SMQT uses an approach that performs an automatic structural breakdown of information. Let x be one pixel and $D(x)$ be a set of $|D(x)| = D$ pixels from a local area in an image. Consider the SMQT transformation of the local area.

$$\text{SMQT (local):} D(x) \rightarrow M(x) \quad (1)$$

These properties are desirable with regard to the formation of the whole intensity image $I(x)$ which is a product of the reflectance $R(x)$ and the illuminance $E(x)$. Additionally, the influence of the camera can be modelled as a gain factor g and a bias term b . Thus, a model of the image can be described by

$$I(x) = gE(x)R(x) + b \quad (2)$$

The SNoW learning architecture is a sparse network of linear units over a feature space [9]. One of the strong properties of SNoW is the possibility to create look-up-tables for classification. Consider a patch W of the SMQT features $M(\mathbf{x})$, then a classifier.

$$\theta = \sum_{x \in W} h_x^{nonface}(M(X)) - \sum_{x \in W} h_x^{face}(M(x)) \quad (3)$$

can be achieved using the nonface table $h_x^{nonface}$, the face table h_x^{face} and defining a threshold for θ . Since both tables work on the same domain, this implies that one single lookup-table.

$$h_x = h_x^{nonface} - h_x^{face} \quad (4)$$

3 Proposed System

This proposed technique was implemented with some attributes derived from an image and the framework is shown in Figure-1. The attributes are derived for every face. In this technique, the faces are treated as a single class. For every single class, six attributes are derived. These attributes have been derived for 4 different images taken in four different days. The attributes are length and height of the face, position of the face in co ordinates, and the number of horizontal lines and vertical lines that are passed from a single class. The framework of the proposed system is shown in the figure 1. In the first step the a sample image, figure 2 is given as an input to the face detection algorithm [2] whose result is shown in figure 3, here some of the faces that are not detected are the ones that are occluded [4], this is shown in figure 4. To detect the occluded faces, the above obtained result is given as an input to the occlusion detection algorithm procedure. The output of the occlusion detection algorithm has been shown in figure 5. Here the overview of the database used in this technique is explained.

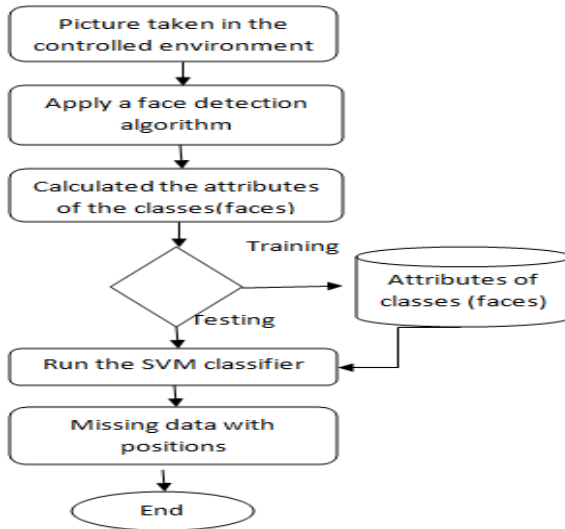


Fig. 1. The Framework of the proposed system



Fig. 2. A Sample image in a dataset



Fig. 3. The output of the detection algorithm

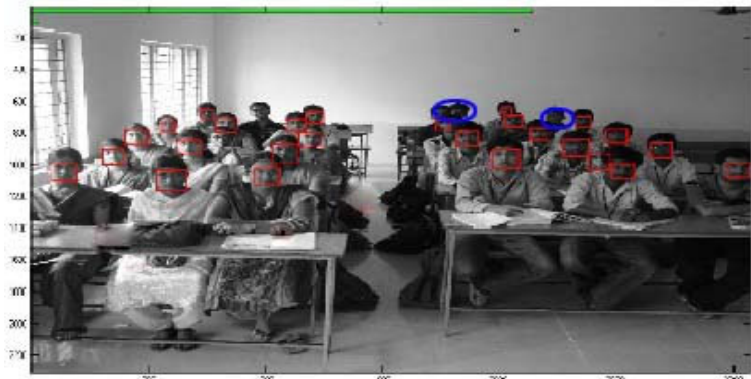


Fig. 4. The occluded faces are highlighted and shown

The dataset of four different days of 30 different classes are updated and maintained in the database followed by the training phase those attributes are used for finding the results. In this technique a closed circuit camera is used to take the pictures. The lighting conditions should be constant.

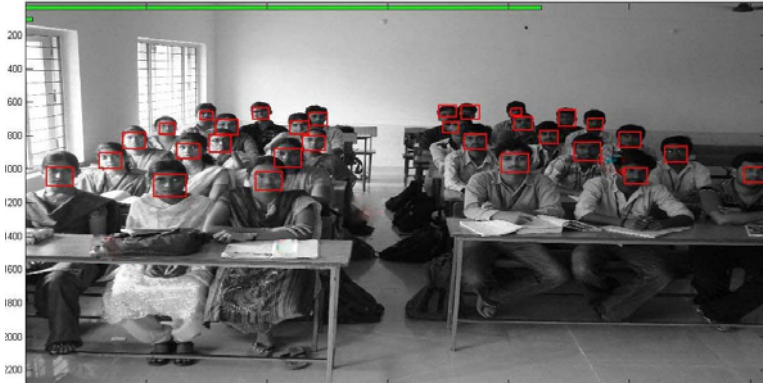


Fig. 5. The result of the Occlusion Detection

Now the process is explained in terms of consecutive steps. In the first step, the image of the classroom is taken and given as an input to the face detection for detecting the faces and creating the bounding boxes around the faces. The sample training image is shown in the figure 2. The training is given with four different images taken in four different days and a single sample image is shown in figure 2.

3.1 First Step

In the first step, a line equation [13] is calculated by using all the points which are selected from all the four training sets of images. The line equation is drawn separately for the rows and columns. A set of five points are taken from every single column of every image and with the help of those 20 points that were selected from those five images, a slope is calculated. The slope is used to plot a curve [13] on the column of every image and this curve will pass almost all the faces in a column. This line is plotted on the first column and the same process and is used to plot all the curves on all columns of an image. The same process is applied for all the rows in the image to draw the curves on all the rows. Therefore a grid is created on the image. The output of the first step is shown in step 4. A grid is formed on all the train images. Two attributes are derived from all the train images with the help of the grid on it. The attributes are the number of horizontal and vertical lines that pass through the faces. The data is derived for all the thirty classes in all the four training images. This will be used for the training stage. Proceeding to the second step, two attributes are updated in the database. The below equations are used to find the best curve which passes through all the faces in the rows and columns.

The equation 5 is used for finding the curve fitting for all the points which are selected in this step. The output along with the grid is shown in figure 6. The below line equation has been used to find the best fitting curve which passes through all the

faces in a row and as well as the columns. The line equation $p(x)$ determines the best fit curve by using all the coordinates of faces in the row and column.

$$P(x) = p_1x^n + p_2x^{n-1} + p_2x^{n-2} + p_3x^{n-3} + \dots + p_nx + p_{n+1} \tag{5}$$



Fig. 6. A grid on the image with horizontal and vertical lines

3.2 Second Step

The second step is used to find the length and height [13] of the faces. Lengths and heights are calculated for all the classes in the image. The calculated lengths and heights of all the faces are updated in the database for training. The respective process is now used to find the lengths and heights of all the four training images. The below geometrical equation is used to find the length and height of each and every class.

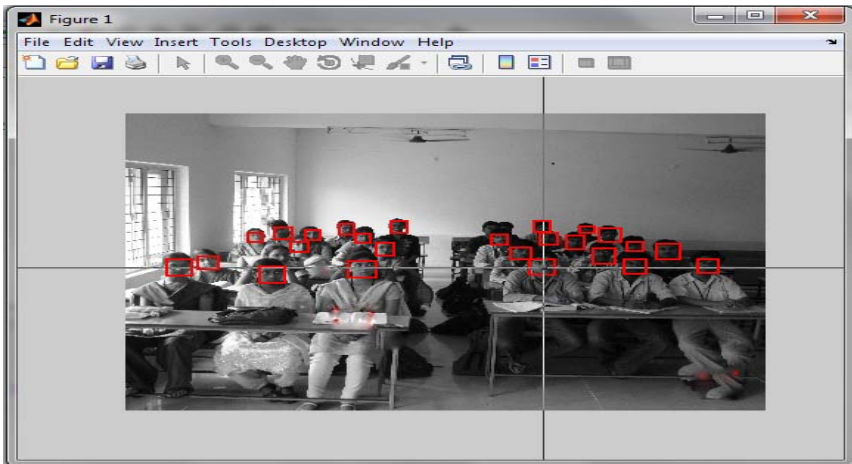


Fig. 7. A snapshot to show how the coordinates are selected

In the second step, all the co ordinates [13] are selected for calculating the length and height of every class. This process has been shown in the figure 7. The coordinates need to be selected for calculating the lengths and heights of the classes. Firstly, all the x coordinates are selected followed by the selection of y coordinates and finally, the z coordinates are selected. Now, with the help of the equations 6 and 7, the lengths and heights of all the classes are derived and updated. The lengths (between x and y), heights (between x and z), positions coordinates, and the number of horizontal lines and vertical lines are passed through the single class. The mathematical equations for calculating the length and height of the class are shown. The respective points are selected from the classes. Three different points are picked from the three different vertex points. The equation 6 and 7 has been used find the length and height of the classes respectively. The equations are derived from the basic co ordinate geometry. The equations are listed below and numbered.

$$L=\sqrt{((x(i+1)-x(i))^2+(y(i+1)-y(i))^2)} \quad (6)$$

$$H=\sqrt{((x(j+1)-x(j))^2+(y(j+1)-y(j))^2)} \quad (7)$$

3.3 SVM Classifier

SVM is a classifier derived from statistical learning theory by Vapnik and Chervonenkis [12]. SVMs are introduced by Boser, Guyon and Vapnik in COLT-92. It was initially popularized in the NIPS community but now is an important and an active field of every Machine Learning research.

Main features:

- By using the kernel trick, data is mapped onto a high-dimensional feature space without much of the computational efforts;
- Maximizing the margin achieves better generalization performance;

SVM requires that each data instance is represented as a vector of real numbers. Hence, if there are categorical attributes, we have to convert them into numeric data. We recommend using m numbers to represent an m-category attribute. Only one of the m numbers is one, and others are zero. For example, a three- category attribute such as red, green, blue can be represented as (0,0,1), (0,1,0), and (1,0,0). Scaling before applying SVM is very important. The main advantage of scaling is to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges. Another advantage is to avoid numerical difficulties during the calculation.

4 Experimentations and Results

This technique is implemented in MATLAB and some part of the technique using the c language. This new technique was implemented and run over a classroom database. The database consists of classroom images of 20 different days. The total classes (faces) present in a single image are 30. In the entire 20 days database, the positions of the thirty classes are fixed and. And a graphical representation of an accuracy of

Table 1. Comparisons of Face Detection algorithms

Number of faces = 30	Skin color based algorithm	Face detection SMQT with split up snow classifier	Face detection with Occlusion detection	Combination of all the algorithm
False Detection Rate (%)	12.5%	16%	11%	3.5%
False Dismissal Rate (%)	20%	24%	8%	2%
Accuracy	67.5	60%	81%	94.5%

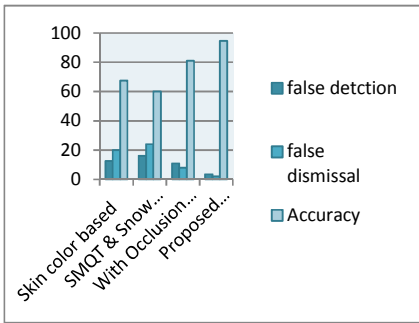


Fig. 8. The Bar Chart to show the comparison of the proposed detection algorithm and the existing algorithms

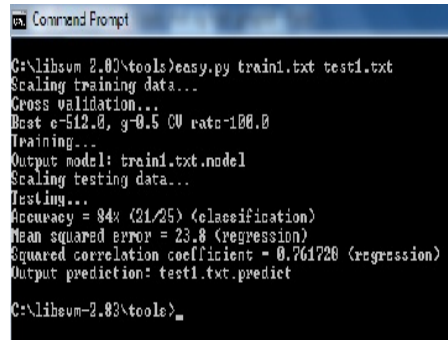


Fig. 9. The training of the SVM



CLASSIFIER OUTPUT

Number of classes found : 26
 Number of missing classes are: 04

Fig. 10. The output of the Classifier missing data was plotted

these algorithms is shown in figure 8. Here the result is created in a text file named predict. This file will give the missing values in the training image. This resultant text file is used to highlight the missing classes in the test image.

Figure 9 shows the accuracy of the System. The missing data is plotted with a blue plot boxes and this has shown in figure 10. The class should be in the intersection of both horizontal and vertical lines, if it is not there then it is missing class, which is marked by the classifier.

5 Conclusions and Future Works

In this paper, a novel technique for detecting missing faces and properly mapping them to specific individuals has been presented. A system based on horizontal and vertical depth lines along with position coordinates has been used as input for SVM Classifier. The results are promising and a good performance is observed in spite of large number of faces and poor illumination conditions.

Also, the applications of the proposed methodology can be extended to various environments where the seating arrangements are fixed like air travel, train travel, seminar halls, laboratories, etc., and advanced curves can be employed for the same.

References

- [1] Turk, M., Pentland, A.: Eigenfaces for Recognition. *Journal of Cognitive Neuroscience* 3(1), 71–86 (1991)
- [2] The Face Detection Homepage, <http://home.t-online.de/home/Robert.Frischholz/index.html>
- [3] Kawaguchi, Y., Shoji, T., Lin, W., Kakusho, K., Minoh, M.: Face Recognition based Attendance System, <http://www.mm.media.kyoto-u.ac.jp>
- [4] Lawrence Zitnick, C., Kanade, T.: A Cooperative Algorithm for Stereo Matching and Occlusion Detection. CMU-RI-TR-99-35
- [5] Vijaya Lakshmi, H.C., Patil Kulakarni, D.: Segmentation algorithm for multiple face detection in color images with skin tone regions using color spaces and edge detection techniques. *International Journal of Computer Theory and Engineering*, 1793–8201 (2010)
- [6] Gonzalez, R.C., Woods, R.E.: *Digital Image Processing*, 2nd edn. Pearson Education, India (2002)
- [7] Nilsson, M., Nordberg, J., Claesson, I.: Face Detection Using Local SMQT Features and Split up SNOW Classifier 20(12), 1222–1239 (2006)
- [8] Lahdenoja, O., Laiho, M., Paasio, A.: Reducing the feature vector length in local binary pattern based face recognition. In: *IEEE International Conference on Image Processing (ICIP)*, vol. 2, pp. 914–917 (September 2005)
- [9] Froba, B., Ernst, A.: Face detection with the modified census transform. In: *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 91–96 (May 2004)

- [10] Nilsson, M., Dahl, M., Claesson, I.: The successive mean quantization transform. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 4, pp. 429–432 (March 2005)
- [11] Yang, M.-H., Kriegman, D., Ahuja, N.: Detecting faces in images: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 24(1), 34–58 (2002)
- [12] Osuna, E., Freund, R., Girosi, F.: Training support vector machines: an application to face detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 1997), pp. 193–199 (1997)
- [13] <http://www.mathopenref.com>

An Experimental Analysis of Phylogenetic Trees Based on Topological Score

Manoj Kumar Gupta, Rajdeep Niyogi, and Manoj Misra

Department of Electronics & Computer Engineering,
Indian Institute of Technology, Roorkee-247667, India
manoj.cst@gmail.com,
{rajdpfec,manojfec}@iitr.ernet.in

Abstract. In this paper we compare phylogenetic trees obtained from two alignment free techniques based on topological score. Whether the constraint followed by their respective graphical representation will generate the practical (appropriate) evolutionary tree. The trees are compared by giving the score based on topological similarity of the branches between the given trees and optimizing the overall score by paring up branches from one tree to another. Phylogenetic trees are constructed by UPGMA, NJ and MEDistance based method. The distance matrix required by tree construction methods are built from two alignment free methods, probability distribution method and genome space construction with biological geometry.

1 Introduction

One of the primary challenges in bioinformatics is dealing with large volume of genomic sequences. A lot of computational and statistical techniques have been proposed by researchers to compare biological sequences. Existing methods for comparing such sequences can be broadly classified into two categories: (i) Alignment based and (ii) Alignment free. Alignment based method requires multiple sequence alignment for many sequences which is computationally difficult. Hence an alignment free approach is desirable. Among existing alignment free methods for comparing biological sequences [1,2,3,4,5,6], graphical representation of a DNA sequence facilitates viewing, sorting and comparing various gene sequences with intuitive pictures and pattern.

Hamori [7] first proposed a 3D graphical representation for DNA sequences, some different graphical approaches representing DNA sequences have been reported by quite a lot of authors. Gates [8] presented the original plot of a DNA sequence as a random walk in a 2D-space using four orthogonal directions to represent the four bases, thereafter Leong and Morgenthaler [9] modify independently. The idea is to read a DNA sequence base by base and plot succeeding points on the graph with four orthogonal unit vectors representing four kinds of bases. The different graphical representation needs numerical characterization to identify the regions of biological interest. Suitable mathematical descriptor helps in finding the numerical (quantitative) representation of similarity or dissimilarity between the sequences, and gives matrix for set of sequences. The elements of similarity/dissimilarity matrix are used to

construct phylogenetic tree. Some means for comparing phylogenies are desirable in order to assess the quality of phylogenetic trees from different construction methods and that are capable of enlightening where two trees agree or differ.

This paper compares two alignment free approaches that use different graphical representations and numerical characterization for measuring similarity between sequences. The trees obtained from three distance based phylogenetic reconstruction methods are compared using Nye et al. [10] method.

2 Background

We have taken two different alignment free techniques proposed by Yu et.al.[11, 12] for obtaining the evolutionary (phylogenetic)relationship among various species. We discuss these methods in following sub sections. Both the methods have been executed on two nucleotide sequence datasets obtained from Genbank. Three distance based method for phylogenetic reconstruction are constructed viz. Unweighted Pair Group Method with Arithmetic Mean(UPGMA)[13], Neighbor joining (NJ) [14] and Minimum Evolution (ME) [15] are obtained from each of the two DNA sequence datasets. These trees are compared by Tom Nye et. al[10] method which gives overall topological score between two trees. The tree comparing algorithm assumes both trees having same set of species and same number of branches. The algorithm for comparing trees performed in two stages, in first stage a score $s(i, j)$ has been assigned to every pair of edges (i, j) present in two trees T_1 and T_2 respectively that reflects the topological similarity of two branches [15]. In second stage branches in two trees are paired up to optimize the overall score. More properly, this is equivalent to finding a bijection (i.e. one to-one correspondence) between the branches of two respective trees that maximizes the quantity [10]

$$\sum_{i \in T_1} s(i, f(i)) \tag{1}$$

where $f: T_1 \rightarrow T_2$ is the bijection function

2.1 Probabilistic Method

Yu et al. [11] make use of probability distribution for DNA sequence and then kullback-Leibler divergence [16] to perform similarity studies. New graphical representation of DNA sequences nucleotides has been used in the probabilistic distribution method. In this representation the four nucleotides are assigned in the first quadrant of the Cartesian coordinate system as shown in figure 1. Figure shows the four vectors corresponding to the four nucleotides A, G, C, and T are as follows: A (1, 0.8), G (1, 0.6), C (1, 0.4), T(1,0.2). The coordinates for four nucleotides confined to first quadrant helps in constructing the probability distribution for DNA sequence. In [11] probability distribution of DNA sequence of length n is defined as

$$p_i = \frac{x_i - \bar{y}_i}{\frac{1}{2}n(n+1) - y_i} \tag{2}$$

where n is the length of sequence x_i and y_i are the coordinate of nucleotide at position i , and \bar{y}_i is the y -coordinate value at the i^{th} nucleotide in the DNA graphical curve according to figure 1.

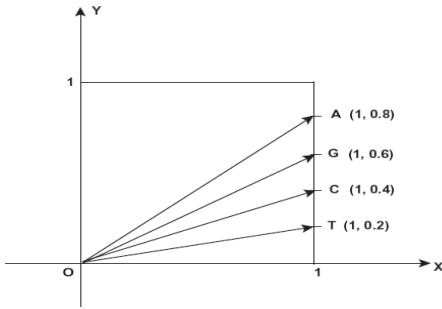


Fig. 1. Nucleotide vector system based on A (1, 0.8), G (1, 0.6), C (1, 0.4), and T (1, 0.2) Source [12]

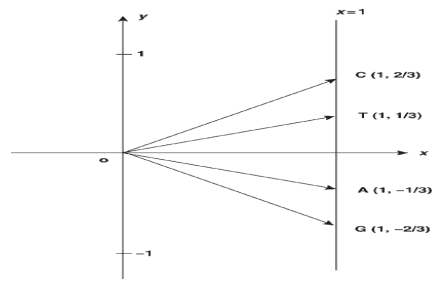


Fig. 2. Nucleotide vector system based on G(1, -2/3), A(1, -1/3), T(1, 1/3), and C(1, 2/3) Source [13]

Since the sequences are of different length, therefore it is transformed to normalize probability distribution by some specific $N < n$. We have taken N as the length of the smallest sequence in the given DNA sequence set. After obtaining the normalized probability distribution for DNA sequences, a similarity/dissimilarity measure between two discrete probability distributions $P_1 = (p_1, p_2, \dots, p_n)$ and $P_2 = (q_1, q_2, \dots, q_n)$ is calculated using Kullback–Leibler divergence (a dissimilarity measure), denoted as $H(P_1, P_2)$ of P_1 with respect to P_2 is defined in [12] as

$$H(P_1, P_2) = \sum_{x \in X} p_1(x) \log \frac{p_1(x)}{p_2(x)} \tag{3}$$

Finally a distance matrix is attained showing the distance between every pair of DNA sequence in the given set of DNA sequences. All coding is done in MATLAB. Following the distance matrix, phylogenetic tree is constructed by UPGMA, NJ and ME distance based method of MEGA 5 package [17].

2.2 Genome Space Construction with Biological Geometry

The other method that we consider here to compute the distance matrix use the idea of genome space with biological geometry. Yu et al.[12] constructed a new DNA sequence graph in two quadrants of the Cartesian coordinate system, with pyrimidines (C and T) in the first quadrant and purines (A and G) in the fourth quadrant as shown in figure 2, the vector corresponding to four nucleotide are as follows: G(1,-2/3), A(1,-1/3), T(1,1/3), and C(1,2/3). The specific ordering of the four nucleotides in the Cartesian coordinate system is related to the GC content of genomes.

Because the GC content of DNA molecule is found to be variable with different organisms, it is considered while assigning the y-coordinate values of nucleotide vectors. Mostly the DNA sequence having low GC content 30–50%, are selected for applying this technique, thus larger y-coordinate absolute values for G and C. However, the y-coordinate values of the four nucleotides must be between 1 to -1 to assure that the correspondence between a DNA sequence and its corresponding moment vector is one-to-one. The advancement of the subject is to construct the moment vectors from DNA sequences using this new graphical method, and these moment vector have one to one correspondence between moment vectors and DNA sequences. The moments for DNA is defined in [12] as

$$M_i = \sum_{i=1}^n \frac{(x_i - y_i)^j}{n^j}, \quad j = 1, 2, \dots, n, \quad (4)$$

where n is the DNA nucleotide sequence length and (x_i, y_i) represents the coordinate position of ith nucleotide in the DNA graphical curve according to figure 2.

The uniqueness of this approach is that by using these moment vectors of DNA sequences, a genome space is constructed as a subspace in Euclidean space. Each genome sequence can be represented as a point in this space. Therefore, this genome space can be used to make comparative analysis to study the clustering and phylogenetic relationship among genomes. The biological (evolutionary) distance between two genomes can be obtained through the Euclidean distance among the corresponding points in the genome space. After obtaining the similar distance matrix as in first method for every pair of DNA sequence we construct the phylogenetic trees by UPGMA, NJ and ME methods of MEGA 5 package [17].

2.3 Data Set

The two methods are tested on two different types of datasets. The graphical representation of four nucleotides in both methods depends on AG and GC content respectively. We selected the first data set (table 1) having AG content of 50.45% and second data set (table 2) having 44.05% AG content. This is appropriate for first method as it requires the low AG content (40%-50%). The GC contents for Table 1 and Table 2 are 56.3% and 44.37% respectively. Exceeding the GC content than

Table 1. Dataset1: Ten beta-globin genes from different species from Genbank

Accession No.	Description	Accession No.	Description
U01317	Human beta globin region	Y00347	European hare beta-globin gene.
AY279114	Woolly monkey beta globin gene	NM_001081704	gallus hemoglobin, beta.
AY279115	Tufted monkey beta globin gene.	X15739	Duck rearranged beta-globin gene
X06701	Rat beta-globin gene.	J03642	Opossum beta-hemoglobin epsilon-M gene
V00882	Rabbit gene for beta-globin.	NM_001123672	Atlantic salmon beta-globin

Table 2. Dataset2: Mitochondrial genome complete sequences of thirty two mammal species

Accession No	Length (nt)	Description	Accession No.	Length (nt)	Description
V00662	16569	H.sapiens	AY488491	16355	Buffalo .
D38116	16563	pygmy chimpanzee	EU442884	16774	Wolf.
D38113	16554	Chimpanzee	EF551003	16990	Tiger .
D38114	16364	Gorilla	EF551002	16954	Leopard
X99256	16472	Common gibbon	X97336	16829	Indian rhinoceros
Y18001	16521	hamadryasbaboon	Y07726	16832	White rhinoceros
AY863426	16389	Vervet monkey	X63726	16826	harbor seal
NC_002764	16586	Barbary ape	X72004	16797	gray seal
D38115	16389	Bornean orangutan	AJ224821	16866	African elephant
NC_002083	16499	Sumatran orangutan.	DQ316068	16902	Asiatic elephant
U20753	17009	Cat.	DQ402478	16868	black bear
U96639	17009	Dog.	AF303110	17020	brown bear
AJ002189	16680	Pig	AF303111	17017	polar bear
AF010406	16616	Sheep	EF212882	16805	giant panda
AF533441	16640	Goat	AJ001588	17245	rabbit.
V00654	16338	Cow.	X88898	17447	hedgehog

constraint for geometry method leads to another conclusion. Table 1 shows the accession number with description of beta-globin genes from 10 different species, these DNA sequences have 444 nucleotides. Table 2 shows the accession number with their description of 32 complete mitochondrial genome sequences each of which has length of more than 16000 nucleotides. All these DNA sequences are from the GenBank (<http://ncbi.nlm.nih.gov/genbank>).

3 Experimental Results

In this section we report the experiments carried out to compare phylogenetic trees based on topological score. We have use two alignment free methods in [11, 12] as given in section 2. All experiments have been performed on our Intel(R) Core(TM)2 Duo CPU E7500 @2.93GHz 2.93GHz, with 2.99 GB of RAM machine in Windows environment. The code for obtaining the distance matrix is executed in MATLAB. Probabilistic distribution method [11] and genome space construction with biological geometry method [12] is tested with beta-globin genes of 10 species and complete mitochondrial genomes of 32 mammals' species.

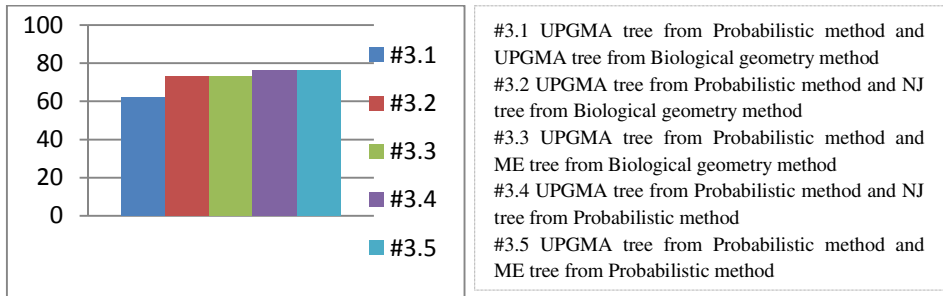


Fig. 3. Topological score by comparing with UPGMA tree from Probabilistic method on dataset 1 given in Table 1

Figure 3 and 4 shows the results for dataset 1. Figure 5 and figure 6 shows the results for dataset 2. The topological score calculated for comparing the phylogenetic tree shows the degree of similarity. High score means higher similarity between two trees. Figure 3 shows the topological scores by comparing the UPGMA[13] tree from probabilistic method[11] to five other trees, viz. three of which are UPGMA, NJ[14] and ME[15] tree from Biological geometry method[12], and two are NJ and ME tree from Probabilistic method. Figure 4 shows the topological scores by comparing the NJ phylogenetic from probabilistic method with four other tree viz. NJ, UPGMA and ME tree from Biological geometry method and ME tree from probabilistic method.

Figure 5 shows the similar comparison of UPGMA tree from probabilistic method with other five trees as in figure 3 but performed on mitochondrial genome sequences of table 2. Similarly Figure 6 shows the topological score by comparing NJ tree from probabilistic method with other four trees i.e. NJ, UPGMA and ME tree from biological geometry method and ME tree from biological geometry method.

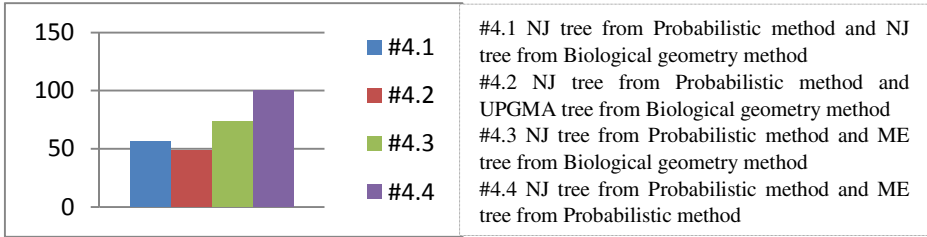


Fig. 4. Topological score by comparing with NJ tree get from Probabilistic method on dataset 1 given in Table 1

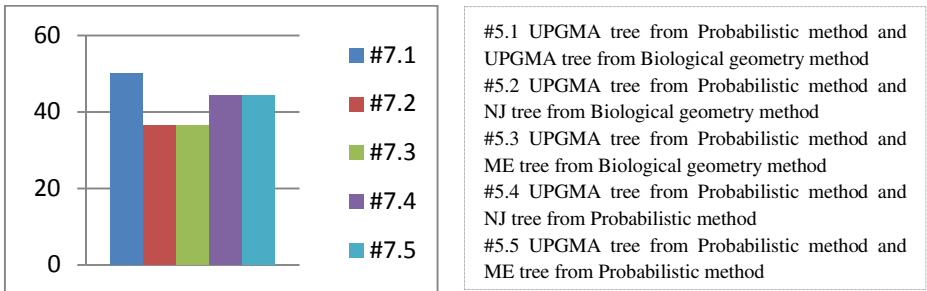


Fig. 5. Topological scores by comparing with UPGMA tree from Probabilistic method on dataset 2 given in Table 2

3.1 Analysis of the Results

From figure 3, we find the lowest score is 62.2% while comparing UPGMA tree from probabilistic distribution method [11] and UPGMA [13] tree from Genome space with biological geometry method [12]. It shows that evolutionary relationship obtained from biological geometry method with UPGMA phylogenetic reconstruction method is not appropriate. The highest score of 76.2% in the same figure 3 is obtained while comparing UPGMA phylogenetic trees and NJ [14] or ME [15] phylogenetic tree from probabilistic method. It shows the evolutionary relationship is approaching the desired one. However, the tree obtained from NJ and ME are quite similar because MEGA employs the Close-Neighbor-Interchange (CNI) algorithm to find the ME tree. NJ or ME tree with probability method compared with UPGMA tree of Biological geometry method gives lowest topological score of 48.5% in figure 4, showing inappropriate phylogeny.

From lowest score 36.5% and highest of 50% in figure 5, we can say that the tree do not give accurate phylogeny by make use of probabilistic method on mitochondrial genome sequence of 32 mammal species as compared with biological geometrical method. In Figure 6 the lowest topological score is 34.1 while comparing NJ tree from both the techniques showing phylogeny is inappropriate.

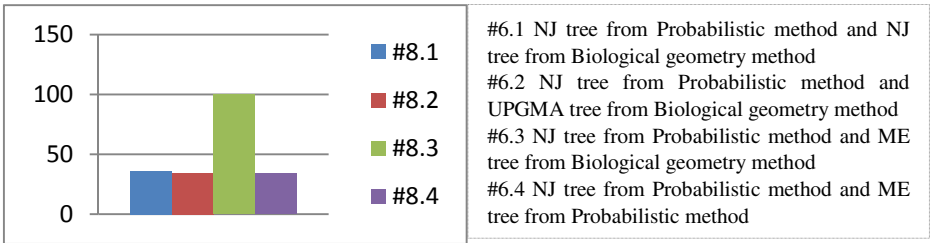


Fig. 6. Topological scores by comparing with NJ tree from Probabilistic method on dataset 2 given in Table 2

3.2 Observation

Figure 7 shows the UPGMA [13] trees obtained from two methods [11, 12]. Tree A is more appropriate than Tree B. In Tree A we find that woolly monkey and capuchin monkey are siblings, i.e., they have a common parent, whereas in Tree B the closest ancestor of woolly monkey and capuchin monkey are at distance of 3 from woolly. The thickness of the branches shows low topological score.

Tree A: UPGMA tree with probability method Tree B: UPGMA tree with Biological Geometry

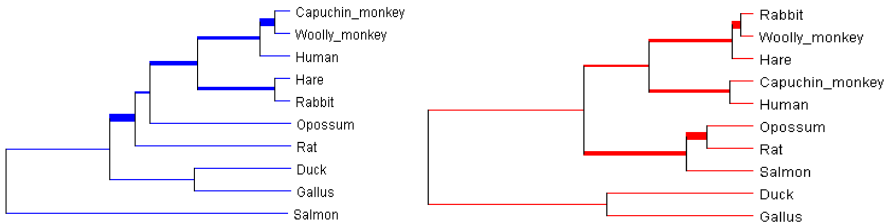


Fig. 7. Two UPGMA trees

Computationally alignment free approaches perform better than alignment for sequences. Moreover, in the smaller length sequences the probabilistic method and geometrical method perform comparative as both takes 44ms and 215ms respectively on beta-globin genes. But for larger sequence length (Mitochondrial genome, Table 2) probabilistic method and biological geometry method takes around 2hrs, 7minutes and 38seconds respectively.

4 Conclusion

We have compared two alignment free DNA sequence comparison techniques; one is based on a probabilistic distribution method and another is Genome space construction with biological geometry. Distance matrices obtained from these two techniques are used for UPGMA, NJ and ME phylogenetic reconstruction method.

These trees are compared with respect to topological score for degree of similarity between trees. Data set is chosen according to the restriction implied by the respective techniques for AG and GC content. Mitochondrial genome sequence follows this restriction but beta-globin genes varies with GC content to 56.3%. Result shows probabilistic method gives more acceptable evolutionary relationship as compared to biological geometry method. We also observe that the graphical representation in biological geometry method of DNA nucleotides is not restricted to have GC content of 40-50% as beta-globin genes have 56.3% GC content. However, the probabilistic method is computationally more intensive than the biological geometric method.

References

- [1] Campello, R.J.G.B., Hruschka, E.R.: On comparing two sequences of numbers and its applications to clustering analysis. *Information Sciences* 179, 1025–1039 (2009)
- [2] Huang, G., Liao, B., Li, Y., Yu, Y.: Similarity studies of DNA sequences based on a new 2D graphical representation. *Biophysical Chemistry* 143, 55–59 (2009)
- [3] Liao, B., Wang, T.: New 2D graphical representation of DNA sequences. *Journal of Computational Chemistry* 25, 1364–1368 (2004)
- [4] Pham, T.D., Zuegg, J.: A probabilistic measure for alignment-free sequence comparison. *Bioinformatics* 20, 3455–3461 (2004)
- [5] Reinert, G., Chew, D., Sun, F., Waterman, M.S.: Alignment-free sequence comparison (I): statistics and power. *Journal of Computational Biology* 16, 1615–1634 (2009)
- [6] Vinga, S., Almeida, J.: Alignment-free sequence comparison – a review. *Bioinformatics* 19, 513–523 (2003)
- [7] Hamori, E., Ruskin, J.: H-curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. *Journal Biological Chemistry* 258, 1318–1327 (1983)
- [8] Gates, M.A.J.: A simple way to look at DNA. *Journal Theoretical Biology* 119, 319–328 (1986)
- [9] Leong, P.M., Morgenthaler, S.: Random walk and gap plots of DNA sequences. *Bioinformatics* 11, 503–507 (1995)
- [10] Nye, T.M.W., Lio, P., Gilks, W.R.: A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. *Bioinformatics* 22, 117–119 (2006)
- [11] Yu, C., Deng, M., Yau, S.S.-T.: DNA sequence comparison by a novel probabilistic method. *Information Sciences* 181, 1484–1492 (2011)
- [12] Yu, C., Liang, Q., Yin, C., He, R.L., Yau, S.S.-T.: A novel construction of genome space with biological geometry. *DNA Research* 17, 155–168 (2010)
- [13] Sokal, R., Michener, C.: A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin* 38, 1409–1438 (1958)
- [14] Saitou, N., Nei, M.: The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4, 406–425 (1987)
- [15] Kumar, S.: A Stepwise Algorithm for Finding Minimum Evolution Trees. *Molecular Biology and Evolution* 13, 584–593 (1996)
- [16] Juang, B.H., Rabiner, L.R.: A probabilistic distance measure for hidden Markov models. *AT& T Technical Journal* 64, 391–408 (1985)
- [17] Tamura, Peterson, Stecher, Nei, Kumar: Phylogenetic and molecular evolutionary analyses were conducted using MEGA version 5 (2011)

A Novel Full Reference Image Quality Index for Color Images

Prateek Gupta, Priyanka Srivastava, Satyam Bhardwaj, and Vikrant Bhateja

Department of Electronics and Communication Engineering,
Shri Ramswaroop Memorial Group of Professional Colleges
Lucknow-227105 (U.P), India
{enggpateek.gupta,priyanka.srivastava4989,
er.satyam123,bhateja.vikrant}@gmail.com

Abstract. This paper presents a new full reference image quality index for objective evaluation of color images in perceptually consistent manner with Human Visual System (HVS) characteristics. The quality index proposed in this paper evaluates the image quality across various types of distortions like: noise contamination, blurring, contrast stretching and compression; by considering the image degradations in terms of error, edge distortion, structural distortion, correlation degree and luminance. Simulation results obtained for the proposed quality index are subjectively validated using Mean Opinion Score (MOS) which ensures the capability and efficiency of the proposed index in evaluating color images according to the HVS characteristics.

Keywords: correlation degree; distortion, full reference, HVS, luminance, MOS, Peak Signal to Noise Ratio (PSNR), quality index.

1 Introduction

Object identification or recognition is very much simplified in color image processing as the color acts as the potent descriptor [1]. Color image processing has therefore become an important part of image processing applications and with this the need of quality evaluation of real time color images is also enhanced. With the ability of human beings to distinguish between numerous color shades in comparison to only few existing gray shades; real time color images can be subjectively evaluated by the human beings in an efficient way. Thus there is a need to develop a quality index for color images which is well correlated with the HVS. Methods for assessment of image quality can be modeled as objective and subjective. Objective evaluation models use the mathematical expressions for assessment of image quality while the subjective models are based on physiological and psychological perception of human. Objective models are used due to their wide acceptability and low complexity. A Full reference image quality index assesses the image fidelity with respect to an ideal reference. These indices are also unaffected from the viewing conditions and individual perception but, sometimes their scores are high even for the images with poor visual quality. Subjective models, evaluate the quality of the image as perceived by an individual observer. However, subjective assessment methodologies are highly dependent on physical conditions and are also inconvenient and time consuming. The

features of both these models are incorporated in the HVS model that correlates well with the perceived image quality. The scope of conventional HVS based evaluation models is limited, as they are dependent upon some basic assumptions; they consider error sensitivity as the only parameter for assessment of image quality [2]. The conventional indexes of quality assessment Mean Squared Error (MSE) and PSNR are operationally simple and have clear physical meaning but are unable to assess the similarity between different distortion types. There are cases when their values saturate to yield similar results with different types of distortions [3]. Structural Similarity (SSIM) [4], models any distortion as a combination of contrast, luminance and structural changes in an image as the human eyes are quite sensitive to the variations in these parameters. Compatibility of SSIM with HVS characteristics accounts for its wide applicability, yet the performance of this metric degrades for poor quality and high texture images [3], [5]. Eric Wharton *et al.* [6] used Weber's Law and Fechner's Law for modeling their metrics. They have also included Michelson contrast and MSE for making their proposed metrics more compatible to HVS. But their metrics do not account for structural and edge changes which are mainly noticed by human eyes. Wei Fu *et al.* proposed a similarity index [7] for color images and used edge, luminance and structural similarity for the assessment of quality. Ho-Sung Han *et al.* [8] used the gradient information for the assessment of image quality. The concept used in their work has been considering only the effect of large differences between the pixels of the original and distorted images. However, distortions projecting small differences in the pixels are neglected. Recently, Chen Yutuo *et al.* proposed an evaluation method for coding color images [9] based on the HVS model. Image pixels, region construction and edges are used as evaluation parameters in their work. The simulations performed by the authors have considered only the limited set of distortions, being silent about the effect of different noise types, rotation, blurring etc. Many objective evaluation methods proposed in the literature have at times proved complex and could not compete over the conventional PSNR [10]-[12]. Although, it is true that images with higher values of PSNR do not often yield good visualization by human observers. This paper proposes a novel full-reference quality index for color images incorporating the known characteristics of the HVS. The proposed quality index is modeled by taking into account the error, structural distortion, edge distortion, correlation degree and luminance. It uses the RGB model for color images and empirically combines the above effects on each of the color plane. Simulations are performed to assess the performance of the proposed method on images affected with different types of distortions like noise contaminations, contrast stretch, blurring, and compression. It has been further validated through subjective evaluation that the proposed quality index is in due coherence with the HVS model. The remaining part of this paper is structured as follows: Section 2 describes the method for estimation of distortions and the proposed image quality index. Details of the simulation procedure, subjective evaluation and the obtained results are discussed under section 3. Finally the conclusions are drawn in section 4.

2 Proposed Evaluation Model

The main purpose of a HVS model is to distillate the structural content from the perception arena. Thus, image degradations can be modeled as losses in the perceived structural content in addition to the error sensitivity estimation (as in Minkowski error

metric [13]). There are still numerous structural changes which cannot be captured by the above estimators; one of them is edge distortion, a primary outcome of the images restored by sharp non-linear filters removing high density noise contaminations at the cost of weak edges and losses of finer details from the image. The motivation for our work is to develop a more precise evaluation framework for degraded colored images by modeling distortion in terms of factors like: error, structural distortion, edge distortion, measurement of correlation degree and luminance for quality evaluation. The procedure for measurement of these distortions is detailed as follows:

2.1 Measurement of Error

The image quality in color image is mainly adjudged by the pixel gray quality of the three color components. The distortion introduced in the image changes the pixel gray value which is mainly noticed by the HVS [9]. Thus, the error introduced in the three color components is calculated separately for quality measurement. If $x(i,j)$ denotes the reference image and $y(i,j)$, the distorted image, then the error introduced in the R component is calculated as:

$$E_r = \frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N [x(i,j) - y(i,j)]^2 \tag{1}$$

where: i & j are the pixel positions of the $M \times N$ image. Similarly, E_g and E_b denotes the error in the G and B components, calculated in a similar fashion as in (1). The overall error introduced in the restored image can be estimated by the average of the error obtained in the three color components. Hence, $PSNR_E$ (dB) for the error-sensitivity in the image is given as:

$$PSNR_E = 10 \log_{10} \left[3 / (E_r + E_g + E_b) \right] \tag{2}$$

2.2 Measurement of Structural Distortion

Human eyes mainly extract the structural changes from the viewing field, so structural distortion is an important parameter for quality assessment. Measurement of this distortion in the proposed work is performed by dividing the image in equal size and non overlapping square regions, along with the calculation of the mean, maximum and minimum pixel values in each region. Structural distortion S_r , S_g and S_b for the three color components R, G and B respectively can be calculated as:

$$S_r = \frac{1}{N} \sum_{i=1}^N \left\{ 0.5 [Xa_i - Ya_i]^2 + 0.25 [Xp_i - Yp_i]^2 + 0.25 [Xb_i - Yb_i]^2 \right\} \tag{3}$$

where: Xa_i , Xp_i , Xb_i and Ya_i , Yp_i , Yb_i denote the mean, maximum and minimum pixel values for the reference and the distorted image respectively. N is the number of regions in which the image is divided. The overall structural distortion is the mean of the structural distortion of the three color components and hence $PSNR_S$ (dB) of the structural distortion is given as:

$$PSNR_S = 10 \log_{10} \left[3 / (S_r + S_g + S_b) \right] \tag{4}$$

2.3 Measurement of Edge Distortion

The quality of an image strongly depends upon the local features of the image like edges, textures and flats. A distorted image with very similar edges to the reference image gives a very high perceptual quality for HVS, although PSNR and SSIM produce an opposite result [7]. Generation of an accurate edge map image is a fundamental pre-requisite in estimating edge distortion. Canny edge detection algorithm [14] is used in this work as it is well-known for its large values of signal-to-noise ratio and high precision. The edge distortion for R component is given as:

$$ED_r = \frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N [x_e(i, j) - y_e(i, j)]^2 \tag{5}$$

where: $x_e(i, j)$ and $y_e(i, j)$ denotes the original and distorted edge maps of the image. In a similar way, the edge distortion can be computed for the G and B components respectively, which are represented by ED_g and ED_b respectively. Hence the overall $PSNR_{ED}$ (dB) of the edge distortion can be given as:

$$PSNR_{ED} = 10 \log_{10} \left[3 / (ED_r + ED_g + ED_b) \right] \tag{6}$$

2.4 Measurement of Correlation Degree and Luminance

Correlation degree [15] between the two images has been calculated in this work to estimate the degree of similarity between the reference and the distorted images. The correlation degree for R component is calculated as:

$$s(x, y)_r = \frac{\sigma_{rx y}}{\sigma_{rx} \sigma_{ry}} \tag{7}$$

In a similar way the correlation degree for G and B component can be calculated and is represented by $s(x, y)_g$ and $s(x, y)_b$ respectively. The overall correlation degree for the color image is given as:

$$s(x, y) = \left(s(x, y)_r + s(x, y)_g + s(x, y)_b \right) / 3 \tag{8}$$

This helps in estimating the actual visual quality of the image and determining level of degradation caused in the image due to the distortions. Luminance [15] also plays as a key role in deciding the visual appearance of the image. Since, the proposed index deals with the HVS characteristics, the effect of luminance change has to be taken in account in order to predict the perceived quality of the image. The luminance for R component is calculated as:

$$l(x, y)_r = \frac{2\mu_{rx}\mu_{ry}}{\mu_{rx}^2 + \mu_{ry}^2} \tag{9}$$

In a similar way the luminance for G and B component can be calculated and is represented by $l(x, y)_g$ and $l(x, y)_b$ respectively. The overall luminance for the color image is given as:

$$l(x, y) = \left(l(x, y)_r + l(x, y)_g + l(x, y)_b \right) / 3 \tag{10}$$

2.5 Proposed Image Quality Index

The proposed image quality index (Q_P) is based on full reference image quality evaluation. It takes the advantages of known characteristics of HVS and has been formulated as the empirical combination of three quality estimation factors i.e. correlation degree, luminance and Q . This can be mathematically defined as:

$$Q_P = s(x, y)I(x, y)Q \quad (11)$$

where:
$$Q = 1 - (1/10^{0.11\alpha PSNR_E + \beta PSNR_{ED} + \gamma PSNR_S}) \quad (12)$$

Q in (12) is the quality factor calculated from the weighted sum of different $PSNR$ (in dB) components based on error, structural and edge distortions obtained from (2), (4) and (6) respectively. α , β , and γ in (12) are the coefficients that have different values empirically determined for different types of distortions. Introduction of different coefficients in (12) is being done as each type of distortion has a different effect on error, structural and edge distortion respectively. In order to make the proposed metric distortion specific, distortions are divided into three cases. Distortions specific to noise attacks are grouped under case A (i.e. Salt & Pepper, Gaussian and Speckle noise). Contrast distortions and degradations due to rotation are considered under the case B whereas case C groups the blurring and compression artifacts. The quality factor (Q) for each case of distortions can be therefore defined as per the given matrix:

$$\begin{bmatrix} Q_A \\ Q_B \\ Q_C \end{bmatrix} = \begin{bmatrix} 0.25 & 0.5 & 0.25 \\ 0.35 & 0.4 & 0.25 \\ 0.35 & 0.25 & 0.4 \end{bmatrix} \begin{bmatrix} PSNR_E \\ PSNR_{ED} \\ PSNR_S \end{bmatrix} \quad (13)$$

where: Q_A , Q_B and Q_C are the quality factor for the distortion cases-A, B and C respectively.



Fig. 1. Reference image simulated with different types of distortion. (a), (b) Salt & Pepper noise. (c), (d) Speckle noise. (e), (f) Gaussian noise. (g), (h) Contrast stretching. (i), (j) Gaussian blurring. (k), (l) Motion blurring. (m), (n) Compression.

3 Simulation Results and Discussion

3.1 Simulation Procedure

In this work a standard Lena image of size 256 x 256 is used as a reference image for the experiments. The early transformations (like normalization, gray conversion etc.) are applied to the input image during the initial pre-processing. Thereafter, different types of distortions are superimposed on the pre-processed image as shown in figure 1. The color image is then decomposed into its three components (R, G & B color planes) for further processing. The *PSNR* component for error sensitivity, structural and edge distortions are then calculated for all the components of the image as described in (1)-(6). Correlation degree and luminance factors are also calculated for each component of the distorted images. Finally, the obtained results are empirically combined to yield the quality index (Q_p) as in (11)-(13), which specifically quantifies the effect of different types of simulated distortion.

3.2 Subjective Evaluation

In this paper the two state-of-art objective indices i.e. PSNR and SSIM along with the proposed index Q_p are tested against subjective evaluation to determine their degree of correlation with the perceived image quality by human observers. In this process 42 images with varying level of distortions like noise contamination, blurring, compression and contrast stretching, etc., were evaluated by a group of observers. These observers rated each image in terms of percentage of perceived quality on a scale between 0 and 1 (0: worst and 1: best). This assessment was performed by a group of 150 observers under strict physical environment that is necessary for proper assessment of the image quality. These evaluations resulted in Mean Opinion Score (MOS), which are the scores given by the individual observers to the images according to the perceived quality. The MOS signifies the rating of the images according to the HVS characteristics and hence gives a basis for subjective quality assessment.

3.3 Comparison of Results

In order to evaluate the performance of the proposed quality metric (Q_p), Pearson correlation coefficient is used as a tool to estimate the correlation of Q_p with respect to subjective evaluation results (MOS) and different HVS based metrics like Edge Performance Index (EPI), Structural Correlation (SC), Mean Absolute Error (MAE) [16] and CONTRAST ($c(x,y)$) [17]. From table 1 it can be observed that the correlation coefficient of Q_p with the MOS is higher in comparison to PSNR and SSIM, showing that the proposed index is more correlated with the HVS and evaluates the image quality in a manner similar to as it is perceived by the human eyes. Thus, Q_p seems out to be a better HVS based metric than conventional PSNR and SSIM. It can also be seen from figure 2(a) that the value of Q_p increases with the increase in the MOS values which shows that the images with higher MOS values

Table 1. Pearson Correlation between different quality measures and subjective evaluation

Distortion Types	PSNR	SSIM	Q_p
Noises	0.9010	0.9439	0.9616
Compressing & Blurring	0.9509	0.9289	0.9442
Contrast	0.8224	0.7571	0.8510

Table 2. Pearson Correlation between different quality measures and EPI and SC

Distortion Types	EPI			SC		
	PSNR	SSIM	Q_p	PSNR	SSIM	Q_p
Noises	0.7802	0.8643	0.9026	0.9769	0.9849	0.9796
Compressing & blurring	0.7776	0.7401	0.7698	0.9998	0.9953	0.9993
Contrast	0.9795	0.9955	0.9680	0.9986	0.9927	0.9976

Table 3. Pearson Correlation between different quality measures and MAE and CONTRAST (c(x,y))

Distortion Types	MAE			CONTRAST		
	PSNR	SSIM	Q_p	PSNR	SSIM	Q_p
Noises	0.8831	0.8771	0.8706	0.9928	0.9610	0.9357
Compressing & blurring	0.9901	0.9759	0.9891	0.8658	0.9090	0.8606
Contrast	0.9583	0.9836	0.9422	0.9775	0.9946	0.9654

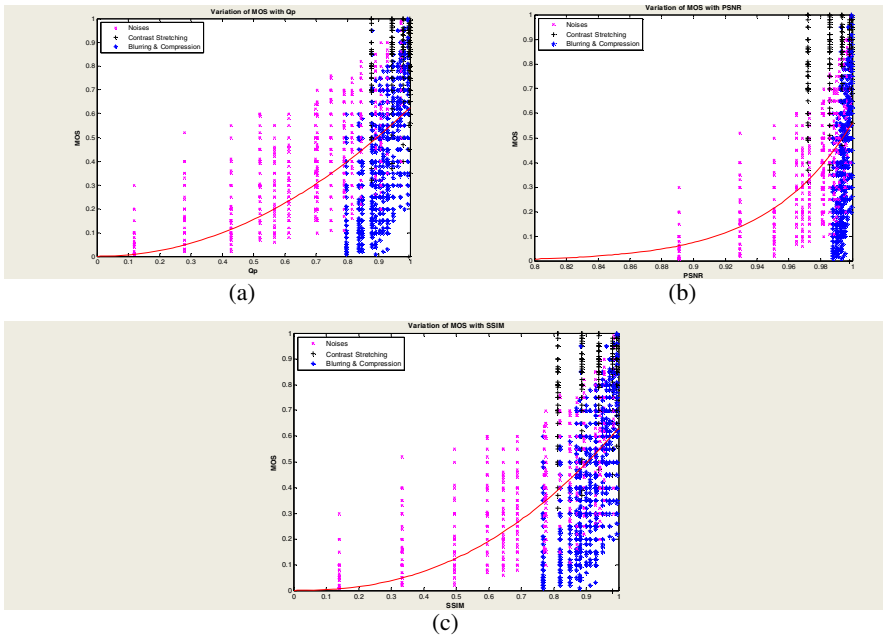


Fig. 2. Scatter plot of MOS Vs different objective evaluation indices. (a) MOS vs Q_p . (b) MOS vs PSNR. (c) MOS vs. SSIM.

images have higher value of Q_p and thereby possess a good visual quality. Similarly, figure 3(b) & (c) depicts the variation of PSNR and SSIM with the MOS values. It can be interpreted from these plots that PSNR and SSIM possess very high values even for images with low MOS values (i.e. poor visual quality). From the results tabulated in table (2) & (3) it can be inferred that apart from being a better HVS based index Q_p possess the capability to assess the edge, structure, error and contrast changes in the image, which is justified from the values of Pearson correlation. These results validate the effectiveness of proposed index in estimating the degradation in the color image for all considered distortion types.

4 Conclusion

The quality index (Q_p), proposed in the paper bear out to be a better HVS based quality evaluation model. The analysis used for developing Q_p incorporates the factors like error, structural distortion, edge distortion correlation degree and luminance, which severely affect the visual quality of color images. All these factors are calculated by using simple and efficient algorithms. The effectiveness of the proposed metric is justified from the results of Pearson correlation calculated between the objective scores of Q_p and subjective evaluation results i.e. MOS. The simulation results also demonstrate that Q_p out performs the conventional PSNR and SSIM. It also shows strong correlation with other metrics like EPI, SC, MAE etc and hence validates its edge, structure and error measuring capability.

References

1. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 2nd edn. Prentice Hall, Upper Saddle River (2001)
2. Wang, Z., Bovik, A.C., Lu, L.: Why is image quality assessment so difficult. In: Proc. IEEE Int. Conf. Acoust. Speech and Signal Processing, Orlando, vol. 4, pp. 3313–3316 (2002)
3. Wang, Z., Bovik, A.C.: Mean Squared Error: Love It or Leave It? IEEE Signal Processing Magazine, 98–117 (2009)
4. Wang, Z., Lu, L., Bovik, A.C.: Video quality assessment based on structural distortion measurement. Signal Process. Image Communication 19(2), 121–132 (2004)
5. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image Quality Assessment: From Error Measurement to Structural Similarity. IEEE Transaction Image Processing 13(4), 600–612 (2004)
6. Wharton, E., Panetta, K., Agatan, S.: Human Visual System Based Similarity metrics. In: Proc. of IEEE Conference on System, Man and Cybernetics, pp. 685–690 (2008)
7. Fu, W., Gu, X., Wang, Y.: Image Quality Assessment Using Edge and Contrast Similarity. In: Proc. of International Joint Conference on Neural Networks, pp. 852–855 (2008)
8. Han, H.S., Kim, D.O., Park, R.H.: Gradient Information- Based Image Quality Metric. In: Consumer Electronics (ICCE), 2010 Digest of Technical Papers, pp. 361–362 (2010)
9. Yutuo, C., Meijie, W., Yongchao, F.: Evaluation Method of Color Image Coding Quality integrating Visual Characteristics of Human Eye. In: Proc. of 2nd International Conference on Education Technology and Computer, Shanghai, China, vol. 2, pp. 562–566 (2010)

10. VQEG.: Final report from the video quality experts group on the validation of objective models of video quality assessment (March 2000), <http://www.vqeg.org/>
11. Martens, J.B., Meesters, L.: Image dissimilarity. *Signal Processing* 70, 155–176 (1998)
12. Eskicioglu, A.M., Fisher, P.S.: Image quality measures and their performance. *IEEE Trans. Communications* 43, 2959–2965 (1995)
13. Pappas, T.N., Safranek, R.J.: Perceptual criteria for image quality evaluation. In: Bovik, A. (ed.) *Handbook of Image and Video Proc.* Academic Press (2000)
14. Canny, J.: A computational approach to edge detection. *IEEE Trans. Pattern Analysis and Machine Intelligence* 8(6), 679–698 (1986)
15. Wang, Z., Bovik, A.C.: A universal image quality index. *IEEE Signal Processing Lett.* 9(3), 81–84 (2002)
16. Tripathi, N., Wahidi, M.F., Gupta, A., Bhateja, V.: A Novel Spatial Domain Image Quality Metric. In: Proc. of (IEEE) 2011 World Congress on Engg. & Technology (CET-2011), Signal & Information Processing, Shanghai, paper-id: 24571 (in press, 2011)
17. Morrow, W.M., Paranjape, R.B., Rangayyan, R.M., Desautels, J.E.L.: Region Based Contrast Enhancement of Mammograms. *IEEE Transactions on Medical Imaging* 11(3), 392–406 (1992)

An Energy Efficient On-Demand Routing by Avoiding Voids in Wireless Sensor Network

J.D. Preethi and R. Sumathi

Department of Computer Science and Engineering,
Siddaganga Institute of Technology, Tumkur
Preethijd03@gmail.com, rsumathi@sit.ac.in

Abstract. Communication voids (i.e. holes) have negative effect on real time routing protocols. To decrease the energy consumption and also avoiding the voids in the sensor networks for real time application it is necessary to design an energy efficient routing protocol which handles holes in the network. In this paper, a new on-demand routing with void avoidance (ODVA) routing protocol has been proposed which efficiently handles and successfully delivers the packets. This protocol also ensures the effective load balancing and improves the network lifetime. The simulation results demonstrate that the ODVA protocol achieve higher packet delivery ratio, network lifetime and efficient usage of energy in comparison with other prevalent scheme.

Keywords: Wireless Sensor networks, Communication voids, Neighbor node information, In-degree, Out-degree.

1 Introduction

Wireless sensor networks (WSNs) are formed by small devices communicating over wireless links without using a fixed networked infrastructure [1]. Because of limited transmission range, communication between any two devices requires collaborating intermediate forwarding nodes, i.e. devices act as routers. Routing in WSNs becomes difficult because there is no infrastructure, wireless links are unreliable, sensor nodes may fail, and routing protocols have to meet strict energy saving requirements. Because of this, many routing algorithm were developed for wireless network. Routing algorithms that perform an end-to-end message delivery with host-based addressing and it can be classified as topology-based, if the destination is given by an ID, and position-based, if the destination is a geographic location. The latter are also called geographic routing. Geographic routing [2] for WSNs has been attracting research interest. Most of the existing geographic routing protocols use greedy routing to forward packets from source to destination. It is a simple, efficient and scalable strategy for geographic WSNs where, a source node selects neighboring node which is closest (with respect to Euclidian distance) to the destination as the next hop, until the destination is reached. It performs well in dense networks, than in sparse networks due to communication voids [3].

A communication void is a state where all the neighbor nodes are farther away from the destination than the node holding the current packet. The node where the packet may get stuck is called as a routing holes or void node. For this reason, early real-time routing protocols based on the greedy routing have received much attention and new techniques are proposed to handle voids in WSNs.

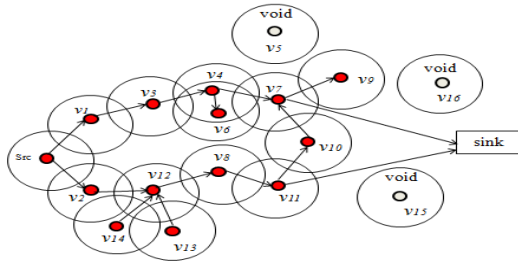


Fig. 1. Illustration of routing holes in WSNs

The scenario of voids in WSNs is as shown in Fig 1; the node v_5 , v_{15} , v_{16} has met the holes in the network because these node cannot find their next hop or neighbor node to reach the destination. Similarly, the node v_{10} can also meets the hole, since its neighbor nodes v_7 and v_{11} are both farther away from the sink node than from itself to the sink node. The node v_{12} can receive the packets from nodes v_2 , v_{14} , v_{13} but it has only one neighbor node to forward the packet further. So the data collision may occur in this node and its battery will be depleted soon.

In this paper, we propose an energy efficient on-demand routing with void-avoidance (OVDA) protocol for sensor networks. The protocol prevents data packets from meeting boundary nodes of voids and balancing load across the network.

The outline of this paper is as follows. Section 2 discusses related work. Section 3, gives an idea of network model, definitions and algorithm description for avoiding the voids, Section 4, shows the results of this methodology and Section 5 discusses the conclusion and advantages of this proposed scheme.

2 Related Work

In this section we summarize the most known protocols in void handling technique.

When a node meets a void, RAP [4] uses GPSR protocol [5], which is based on the right hand rule, to route packets around the perimeter of the void region. However, many packets are dropped because their deadlines are expired in long routes. SPEED [6] and MMSPEED [6] handle voids in the same way as they handle congested regions, where data packets are dropped by stuck nodes and a backpressure beacon is issued to upstream nodes to prevent further drops. However, it has been admitted in these protocols that the void avoidance scheme is not guaranteed to find a path if there is one as in GPSR [5], but it is guaranteed to find a greedy path if one exists. RPAR [7] incorporates a face routing mechanisms to route packets around large voids in

sensor networks. FT-SPEED [8] [9] borrows the idea of SPEED to handle the real time packet delivery in greedy forwarding. Based on the right-hand rule, FT-SPEED proposes an alternative void bypass scheme where data packets at stuck nodes are routed around two sides of the void. At stuck nodes, FT-SPEED will no longer take the real-time requirement as the criteria to choose the next hop node. Also, it uses the void edge nodes to deliver data packets and control packets to maintain the void fresh information, data collisions may occur in these nodes and their battery will be depleted soon.

3 Network Model and Definitions

3.1 WSN Model

WSNs consists of large number of sensor nodes deployed to a vast field with single base station at the center as depicted in Fig 2. Here, network consists of two types of sensors: small number of powerful *Source Nodes* (SNs) responsible for monitoring and initiating data transmission whenever the event is detected and large number of low end *Intermediate Nodes* (INs) responsible for data transmission. Each node has its specific radio range with radius r and can directly communicate with any node within its radio range. In this scenario, we consider event driven wireless sensor application. In other words, when the node has to forward the packet the source initiates the route to reach the destination.

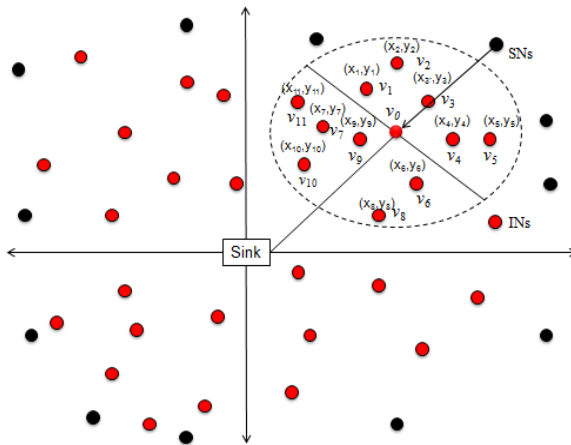


Fig. 2. A WSN model

3.2 Definitions

Neighbor list information. As soon as nodes are deployed in the network, each node start constructs their neighboring node set. Neighbors of a node v_i are defined as

$N(v_i) = \{v_j \in V \mid d \leq r, i \neq j\}$ where, d is 2D-Eculidean distance and is given by $d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$. The neighboring nodes for node v_0 in Fig 2, $N(v_0) = \{v_1, v_2, v_3, \dots, v_{11}\}$. This information is stored in Neighbor Information Table (NIT) as shown in Table 1.

Table 1. Neighbor Information Table

Node_ID	(x ,y)
v_1	(x_1, y_1)
v_2	(x_2, y_2)
v_3	(x_3, y_3)
v_4	(x_4, y_4)
v_5	(x_5, y_5)
v_6	(x_6, y_6)
v_7	(x_7, y_7)
v_8	(x_8, y_8)
v_9	(x_9, y_9)
v_{10}	(x_{10}, y_{10})
v_{11}	(x_{11}, y_{11})

Out-Degree. An out-degree for sensor nodes $v_i \in V$ is the number of downstream nodes which is currently receiving packets to v_i . The downstream of nodes is the flow of nodes which is closer to the sink node. It is represented as given in the equation below,

$$N(v_i) = \{v_j \in N \mid \theta \leq 180^\circ \mid i \in (v_i), i \neq j\} \tag{1}$$

Where, $\theta = \cos^{-1}(a_1 a_2 + b_1 b_2) / \sqrt{(a_1^2 + b_1^2)(a_2^2 + b_2^2)}$, $a_1 = (x_1 - x_2)$, $b_1 = (y_1 - y_2)$, $a_2 = (x_3 - x_4)$, $b_2 = (y_3 - y_4)$ are position of nodes. The out-degree of a node $v_0 = \{v_6, v_7, v_8, v_9, v_{10}, v_{11}\}$ as shown in Fig 2.

In-Degree. An in-degree for sensor nodes $v_i \in V$ is the numbers of upstream nodes which are currently forwarding the packets to v_i . The flow of node which are closer to the source node are called as upstream nodes and it is given by, $v_i(\text{in-degree}) = N(v_i) - v_i(\text{out-degree})$. The in-degree of a node $v_0 = \{v_1, v_2, v_3, v_4, v_5\}$ as shown in Fig 2.

Out-Degree/In-Degree ratio (R). It is the ratio of downstream and upstream nodes for a node v_i in the network. For e.g. as represented in Fig 2, the ratio of node v_0 will be 6/5 which equals to 1.02. The construction of in-degree & out-degree ratio as explained detail in below Algorithm I.

Algorithm I: In-degree and Out-degree ratio construction

Input: Neighbor information table

Output: In-degree and Out-degree ratio for each node

Step 1 Compute the In-degree and out-degree of nodes.

$$\theta = \cos^{-1}(a_1 a_2 + b_1 b_2) / \sqrt{(a_1^2 + b_1^2)(a_2^2 + b_2^2)}$$

if ($(v_i) \leq 180^\circ$) then v_i =out-degree

else

$$v_i \text{ (in-degree)} = N(v_i) - v_i \text{ (out-degree)}$$

Where, $N(v_i)$ all neighboring nodes.

Step 3 Flood the hello packet into network to know the in-degree and out-degree ratio of neighboring node

Step 4 Choose the node to forward the packets

To know the in-degree & out-degree ratio of each neighboring node, exchange the hello packets along with <Node ID, (x, y), R> and this information is updated in the NIT. By using this method it can easily avoid the void in the network. Updated NIT is as shown in Table 2.

Table 2. Updated Neighbor Information Table

Node ID	(x, y)	R
v_1	(x_1, y_1)	4/3
v_2	(x_2, y_2)	3/2
v_3	(x_3, y_3)	2/3
v_4	(x_4, y_4)	SP
v_5	(x_5, y_5)	3/3
v_6	(x_6, y_6)	SP
v_7	(x_7, y_7)	2/2
v_8	(x_8, y_8)	2/4
v_9	(x_9, y_9)	SP
v_{10}	(x_{10}, y_{10})	3/3
v_{11}	(x_{11}, y_{11})	4/2

4 On-Demand Routing with Void Avoidance Routing Protocol

4.1 Protocol Description

ODVA is proposed in order to avoid voids in WSNs. It consists of two phases:

Neighbor Creation. As soon as the nodes are deployed in the network each node starts constructing its NIT as explained in section III. To collect the neighbor node information each node exchanges the hello messages along with information $\langle \text{Node ID}, (x, y), R \rangle$. Upon collecting the neighbor node information, the node chooses next node to forward the packets.

Data Routing Phase. In on-demand routing, each node dynamically constructs its route to reach the destination. When the source initiates the route to forward the packet, first it checks in-degree and out-degree ratio value in NIT. If the ratio value, R , is greater than 1, i.e., if the node has less upstream nodes than the downstream nodes which means no voids occur then choose the node to forward the packet. If $R=1$, then the nodes have equal number of upstream and downstream nodes. In this case, chances of avoiding the voids are more. If $R<1$, then the node has more upstream nodes than downstream nodes. In this case, choose the node having less upstream nodes among the nodes so that the chances of avoiding the voids are very less. If $R=SP$, then the node is having only one upstream and downstream node each. Choosing this node should be avoided since there exists no next hop or neighbor node to forward the packets. By considering the ratio value, each node selects the neighbor node to forward the packet.

To avoid choosing the same node each time, the node uses pointer called router information pointer (RIP). It maintains the information of $\langle \text{Node ID}, \text{downstream nodes}, \text{upstream nodes}, \text{Count} \rangle$ in routing information table as shown in TABLE III. The count of each node increases each time it is chosen to forward the packet. Before choosing the node, it checks for the count and the node with lesser count value is chosen. Algorithm II explains the selection of nodes to forward the packet.

Algorithm II: Selecting the nodes to avoid the voids

Input: Routing information table

Output : Selecting the node to avoid the voids

```

Step 1  begin
        for all  $i=0$  to  $n$ 
        Choose the node with  $R>1$  and count  $c=i$ 
        then choose the node
        else
            go to step 2
Step 2  Choose the node  $R=1$  and count  $c=i$ 
        then choose the node
        else
            go to step 3
Step 3  Choose the node  $R<1$  and count  $c=i$ 

```

- then choose the node
 - Step 4* if $R=SP$,
then avoid to choose the node
 - Step 5* Once all the count value becomes 1
then repeat same process
-

Table 3 maintains the routing information or ratio value of each node. It helps to choose the node by considering the ratio value, to avoid the voids and balancing the load across the network.

Table 3. Routing Information

Node ID	No. of downstream nodes	No. of upstream nodes	count
v_2	3	2	1
v_3	2	3	0
v_4	1	1	SP
v_5	3	3	0

Let the nodes v_2, v_3, v_4 and v_5 be the intermediate nodes. According to the definition of in-degree and out-degree, it is observed that node v_2 has 2 upstream and 3 downstream nodes. As a result the ratio will be $3/2$ which is greater than 1 ($R>1$). Before choosing that node first it checks for the count. The count value of node v_2 is 1 i.e., it is already chosen once. Next node will be the node v_5 which has equal number of upstream and downstream node, i.e. $3/3$ which is equal to 1 ($R=1$). The count of node v_5 is 0, so choose this node to forward the packets. After choosing this node the count becomes 1. The next node to forward the packet is node v_3 which has more upstream nodes than downstream nodes ($R<1$), but the count of node v_3 is 0 so choose this node to forward the packet. After choosing this node the count becomes 1. Once the count value of all nodes becomes 1 then repeats the same process. The node v_4 in Table 3 having ratio $1/1$ which is specified as special node ‘SP’ must be avoided from choosing.

5 Simulation and Results

Extensive simulations are carried out to analyze the effectiveness of the proposed void avoidance mechanism. The simulated network model consists of 100 sensor nodes of which are populated over an area of 100×100 sq. meters with radio range r . The nodes are distributed randomly and uniformly over the deployment area. Each data

packet is about 30 byte and energy per node is 5 joule. Applying the technique described in the above section gives to avoiding void in the network.

Packet delivery ratio. The average packet delivery ratio of ODVA is compared with the protocol FT-SPEED as shown in Fig 3. The packet delivery ratio for ODVA is always same if the voids increase by 10% for 100 nodes and 20% for 200 nodes in the network. However, FT-SPEED decreases the packet delivery ratio. Fig 4 shows that the ODVA protocol gives better average packet delivery ratio than FT-SPEED.

Energy consumption. Energy consumption of ODVA is compared with FT-SPEED as shown in Fig 4. When the transmission range $R=500$, the energy consumed by the ODVA protocol is 1J but in the FT-SPEED it varies. In ODVA even if the range increases, an equal amount of energy is consumed by each node so that it balances the load across the network. Fig. 4 gives energy consumed by the ODVA protocol is less than the FT-SPEED.

Network lifetime. The network lifetime for the ODVA protocol is much higher than FT-SPEED is as shown in Fig 5. We can see that network with FT-SPEED simply runs within the 40 seconds of the simulation when $N=50$ and 35 seconds for $N=100$ while network with ODVA protocol runs 75 seconds when the $N=50$ and 80 seconds of the simulation when $N=100$. These gains are expected to be further magnified in larger networks running for greater intervals of time.

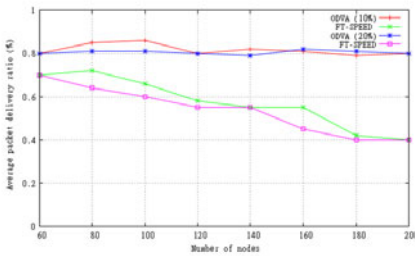


Fig. 3. Average packet delivery ratio

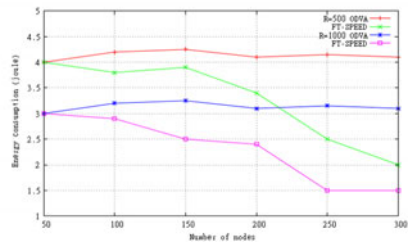


Fig. 4. Energy consumption in network

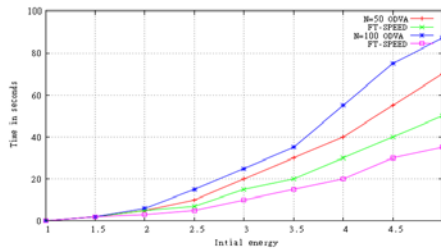


Fig. 5. Network Lifetime

6 Conclusion

In this paper, we proposed an ODVA protocol by introducing in-degree & out-degree ratio at each node. The ratio 'R' value at particular node presents the useful information to select node. The ratio R aims at reducing voids in the network and enables a higher degree of fairness. In the proposed, ODVA protocol each node dynamically constructs its route to balance the load and improve the network life time. This technique yields itself well to achieve efficient routing and avoids the void in the network. When compared with FT-SPEED protocol the proposed ODVA protocol provides the good performance in terms of packet delivery ratio, energy consumption and network lifetime.

References

- [1] Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: A survey on sensor networks. *IEEE Communications Magazine*, 102–114 (2002)
- [2] Capkun, S., Hamdi, M., Hubaux, J.: Gps-free Positioning In Mobile Ad Hoc Networks. In: *Proceedings of Hawaii International Conference on System Sciences*, Hawaii, USA, pp. 1–10 (January 2001)
- [3] Aissani, M., Mellouk, A., Badache, N., Boumaza, M.: A novel approach for void avoidance in wireless sensor networks. *International Journal of Communication Systems* 23, 945–962 (2010)
- [4] Lu, C., Blum, B.M., Abdelzaher, T.F., Stankovic, J.A., He, T.: RAP: A Real-Time Communication Architecture for Large-Scale Wireless Sensor Networks. In: *IEEE Real-Time and Embedded Technology and Applications Symposium, RTAS 2002* (September 2002)
- [5] He, T., Stankovic, J.A., Lu, C., Abdelzaher, T.F.: A Spatiotemporal Communication Protocol for Wireless Sensor Networks. *IEEE Transactions on Parallel and Distributed Systems* 16(10), 995–1006 (2005)
- [6] Felemban, E., Lee, C.G., Ekici, E.: MMSPEED: Multipath Multi- SPEED Protocol for QoS Guarantee of Reliability and Timeliness in Wireless Sensor Networks. *IEEE Transactions on Mobile Computing* 5(6), 738–754 (2006)
- [7] Chipara, O., He, Z., Xing, G., Chen, Q., Wang, X., Lu, C., Stankovic, J.A., Abdelzaher, T.F.: Real-time Power-Aware Routing in Sensor Networks. In: *IEEE International Workshop on Quality of Service* (June 2006)
- [8] Jia, W., Wang, T., Wang, G., Guo, M.: Oriented Void Avoidance Scheme for Real-Time Routing Protocols in Wireless Sensor Networks. In: *Proceedings of the IEEE International Conference on Global Telecommunications*, pp. 1–5 (2008)
- [9] Zhao, L., Kan, B., Xu, Y., Li, X.: FT-SPEED: A Fault Tolerant, Real-Time Routing Protocol for Wireless Sensor Networks. In: *International Conference on Wireless Communications, Networking and Mobile Computing (WiCom)*, pp. 2531–2534 (2007)

GP Boosting Classification on Concept Drifting Data Streams

Dirisala J. Nagendra Kumar¹, J.V.R. Murthy²,
Suresh Chandra Satapathy³, and S.V.V.S.R. Kumar Pullela⁴

¹ BVRICE, Bhimavaram, India

² JNTUCE, Kakinada, India

³ ANITS, Visakhapatnam, India

⁴ VS Lakshmi Engg. College, Kakinada

{nagendrakumardj, mjonnalagedda,
sureshsatapathy, ravipullela}@gmail.com

Abstract. Genetic Programming is an evolutionary soft computing approach. Data streams are the order of the day input sources. In general, data streams exhibit a peculiar behavior of drifting the concepts as time passes by. Here is a study of GP Classifier on Concept Drifting Data Streams. GP classifier performance is compared to that of other state-of-the-art data mining and stream classification approaches. Boosting is a machine learning meta-algorithm for performing supervised learning. A weak learner is defined to be a classifier which is only slightly correlated with the true classification. In contrast, a strong learner is a classifier that is arbitrarily well-correlated with the true classification. Boosting combines a set of weak learners to create a strong learner. It is observed that the Boosting GP approach is beating Boosting Naïve Bayes classification on Concept Drifting Data Streams. Hence it is found that GP is a competent algorithm for Concept Drifting Data Stream classification.

Keywords: Genetic Programming, Classification, Multi-class, Boosting, Data Stream, Stream Mining, Concept Drifting Data Stream.

1 Introduction

Many organizations are being dumped with tremendous amount of continuous flow of data, due to a sequence of events from different locations of the organization. Telephone records, credit/debit card transactions, sensor networks, network event logs, web log data, online sales transactions are some examples of data streams. Traditional approach for mining data is known as batch processing, as it assumes data as a static entity. Now data streams stresses the need of online and incremental data mining techniques, of course should be able to deal with concepts drifts in some cases.

Classification is a major data mining technique [6][13]. Genetic Programming (GP) is one of the famous classification techniques, which has its roots in Genetic Algorithms (GA) [7-9]. Bagging and Boosting are two meta learners in data mining [6][12][13].

Data stream classification is studied in [15-20]. Mohammad M. Masud et. al. studied Mine Class Algorithm for automatic detection of a novel class in presence of concept-drift [15]. Gianluigi Folino et. al. has studied a StreamGP approach with adaptive boosting ensemble algorithm for classifying homogeneous distributed data streams [16]. Hussein A. Abbass et. al. made a detailed study of online adaption in learning classifier systems for stream data mining based on Genetic Algorithms [17]. Yi Zhnag and Xiaoming Jin built an ensemble classification technique on data streams [18]. New ensemble methods for evolving data streams are studied by Albert Bifet et. al.[20]. Wenyan Wu and Le Gruenwald studied various issues involved in simultaneous mining of multiple data streams [19].

Most of the work on classification concentrates on binary classification problems. Traditionally Maximum Likelihood Classifier (MLC) [10], Bayesian networks [10], and Neural networks (NN) [11] are the most successful approaches for multi-class classification.

Genetic Programming (GP) is a stochastic approach, derived from Genetic Algorithms (GA), to solve various computer related problems by automatically constructing programs simulating the biological evolution [8]. GP is a nice approach for solving the binary and multi-class classification problems. It guarantees good classification accuracy if enough training time is given to evolve a higher accuracy GP classifier [2]. An attempt is made to reduce this training time to a reasonable degree. The goals that are tried to meet are simplicity, scalability, and high accuracy. The GP classifier has to find fitness for all fitness cases, which may not be stored in main memory for larger datasets. In order to achieve scalability, the size of training data set sampled at a time is restricted to a portion of main memory available. Topan Kumar Paul, and Hitoshi Iba [5] implemented the ensemble approach of Boosting based on GP and called it “a majority voting genetic programming classifier”.

T. Loveard and V. Ciesielski [1] proposed five alternative methods to perform GP-based multi-class classification, viz., Binary decomposition, Static range selection, dynamic range selection, class enumeration and evidence accumulation.

J. K. Kishore et al.[2] modeled the n-class pattern classification problem as an n two-class problems. A Genetic programming classifier expression (GPCE) is evolved as a discriminant function for each class. Each GPCE recognizes data samples belonging to its own class and rejects samples belonging to other classes. In [2]-[4], [14] the authors designed a classifier with n binary-trees for the n-class classification problem.

D.P. Muni et al. [3] improved the approach of J.K. Kishore et al.[2] by generating the classifier in one pass. D.P. Muni et al. extended their earlier work to suit for feature selection (FS) in [4], proposing a wrapper approach for FS.

Topan Kumar Paul, and Hitoshi Iba [5] proposed a majority voting technique, which evolves multiple GP rules and apply those rules to test samples to determine their labels and count their votes in favor of a particular class. Then the sample is assigned to the class that gets the highest number of votes in favor of it.

T. Loveard and V. Ciesielski[1] used the total training set as exemplar set. In [3], [4], D.P. Muni et al. used step-wise learning, which takes a smaller exemplar set initially, and gradually increases the exemplar set to the whole training set.

2 Data Streams

The recent advances in hardware and software have enabled the capture of various measurements of data in a wide range of fields. These measurements are generated continuously and in a very high fluctuating data rates. Examples include sensor networks, web logs, and computer network traffic. The storage, querying and mining of such data sets are highly computationally challenging tasks. Mining data streams is concerned with extracting knowledge structures represented in models and patterns in non stopping streams of information. The research in data stream mining has gained a high attention due to the importance of its applications and the increasing generation of streaming information. Applications of data stream analysis can vary from critical scientific and astronomical applications to important business and financial ones. Algorithms, systems and frameworks that address streaming challenges have been developed over the past decade. There is a real need inspired by the potential applications in astronomy and scientific laboratories as well as business applications.

3 GP Boosting Approach for Data Stream Classification

Once a GP classification program predicting the class labels is built, it can be used directly for classification, or combine the GP programs into a more efficient solution. There are several ways to combine classifiers. For example, one can use a voting system for the results of several classifiers.

In the case of a classification problem with n classes, there are several approaches to build classifiers. The three most common approaches are:

1. Develop a single classifier that gives, as output, the class of the new sample as input.
2. Develop n classifiers. Each classifier is responsible for recognizing a particular class.
3. Develop a classifier for each pair of classes. Each classifier is responsible to decide between two classes in particular.

The method 2, n -classifier approach, is the best approach [23]. Thus a variant of method 2 is used in this study. With this method, the classifiers obtained by the GP must have some type of output value. Two approaches were again proposed:

1. Binary classifier

If the classifier result is 0, it predicts that the sample is not part of the class, or if 1 predicts that the sample belongs to the class.

2. Classifier with Continuous Output

The result is a decimal value (eg between 0.0 and 1.0) that represents the confidence with which the classifier links the sample with the designated class.

When a new sample is introduced, each classifier must predict whether the sample belongs to the class for which it was trained. The combined classifier has the output value that determines the largest class of the new sample. In the event of a tie, the classifier that has the highest probability will be identified as the class of new sample.

The present work integrates the boosting meta learner with the evolutionary process of GP. Boosting algorithm is applied on n-class GP classifiers. Here each classifier predicts the confidence with which the classifier is assigned the class. Several studies on the implementation of a method for boosting the GP have reported significant gains in terms of classifier accuracy and computation time of the algorithm [21-22]. The integration of the principles of boosting even within the GP process allows greater economy of resources. Here is the pseudo code of the Boosting GP approach adapted here:

```

C = number of classes of the problem
P = number of necessary programs for boosting
Training set, T = all training data available
N = total number of samples in T
For all Ci (j = 1 to C)
    Empty the GP population, POP
    Initialize the weight of each sample W with Wi = 1/N
    For all Pk (k = 1 to P)
        If POP is empty, fill POP with a new set of programs.
        Changing a program that recognizes the class C (the calculation of
        fitness uses the weight Wi of each sample to classify), using T and POP.
        Calculate the error of the best program, Ejk on the training set, a
        factor αjk and then the weight of each sample Wi using AdaBoost method.
    End for P
End for C
    
```

A sample can be classified using the strongest response in a weighted sum of the outputs of programs by class (using equation 1).

$$\max \left(\sum_{k=1}^P (a_{jk} * \alpha_{jk}) \right) \tag{1}$$

By the end of the routine, P*C programs (where P programs for each class of the C classes) are obtained. The classification score for class j is obtained by the weighted sum by α_{jk} output of each program jk. The class that scores the highest indicates the class of the given sample:

4 Fitness Function for GP Boosting Approach

The fitness is the measure of GP program performance in the prediction of output values from input samples. It is therefore an indication of relevance of the program for classifying the samples in training dataset. Fitness is a numeric value, allowing us to compare the performance of programs. Fitness is used to select programs in the population to transform further.

Fitness function is the result of classifications on the training data. This function compares the value of predicted class and actual class provided in the training data. The fitness function depends on the approach used in the combination of classifiers.

1. For a single classifier approach, fitness is simply the number of correct predictions of the program. This value can be normalized (between 0.0 and 1.0) by dividing the number of matching samples in the data set by the total number of samples in training dataset.

2. In the case of an approach of n-class classifier, the calculation of the fitness depends on the classifier chosen:

a) Binary Classifier: It is as in the single classifier approach.

b) Classifier with Continuous Output: The output of the program P is a value of limited trust between -1.0 and 1.0. Fitness is calculated from the sum of S values of confidence of P_i for each sample, depending on the class C provided by the training data set.

$$S = \sum_j P(i) * C(i) \tag{2}$$

C (i) is 1.0 if the sample i belongs to the class recognized and -1.0 otherwise. Finally, fitness is the sum of S values, normalized between 0.0 and 1.0.

c) Classifier output continues to boosting algorithm built: The technique is essentially the same as (b), but the weight W of training samples is taken into account:

$$S = \sum_j P(i) * W(i) * C(i) \tag{3}$$

As the weight of the samples is also normalized (total weight is 1.0), the sum S can be normalized in the same way as in (b). So for a classification problem, the more fit, the more the program is effective. A perfect prediction rate is obtained when the fit is 1.0. Here 2(c) approach is followed. Every Genetic Programming approach needs some parameters to be specified. In this approach, the GP parameters used are given in Table 1.

Table 1. The default GP parameters used for GP Classifier Construction

Parameter	Values
Population size	100
Maximum depth	5
Stopping Criteria	Fitness=99%, Max. Generations=100, Max. Time=5 min.
Population Initialization	Ramped-half-and-half
Selection	Roulette wheel
GP operator proportions	Crossover=90%, Mutation=7%, New Program=3%

5 Results

The data on which the classifiers are executed are 2-class and 5-class Concept Drift Random trees each with 10 Million rows and the evaluation is through interleaved test then train evaluation. The result of GP classification on the above datasets is as follows:

Table 2. The time taken and classifier accuracy % of various classifiers on 2-class Concept Drift Random trees

Classifier	Functions	Time in sec.	Accuracy %
AdaBoost M1+ NB	--	1h4m55s	72.02
GP	+, -, *, /	59m38s	73.01
GP	+, -, *, /, If, <, >	1h4m34s	73.85
GP	If, <, >	31m18s	57.82
GP	If, <, >, !, &,	30m32s	71.17

Table 3. The time taken and classifier accuracy % of various classifiers on 5-class Concept Drift Random trees

Classifier	Functions	Time in sec.	Accuracy %
Adaboost M1+ NB	--	1h40m19s	54.20
GP	+, -, *, /	1h42m37s	54.73
GP	+, -, *, /, If, <, >	2h5m12s	45.31
GP	If, <, >	50m28s	37.54
GP	If, <, >, !, &,	1h20m11s	43.43

In case of the above 2-class Random tree dataset, Boosting GP with +, -, *, /, If, <, > functions classifying with 73.85% accuracy is better than that of the combination of AdaBoostM1 and Naïve Bayes Classification with 72.02% accuracy. And for 5-class Random tree dataset, Boosting GP with functions +, -, *, / classifying with 54.73% accuracy is above that of the combination of AdaBoostM1 and Naïve Bayes classifier with 54.20% accuracy. The only disadvantage is that there is no single combination of GP functions and parameters suitable for all datasets. Hence in general, this Boosting GP is a good candidate for Concept Drifting Stream classification and is suitable for further work.

6 Conclusions

It is found that Boosting GP Classifier is a competent approach for classifying concept drifting data streams. The issue is changing accuracies of GP classifier with functions and GP parameters. The next goal is to improve the GP approach in two respects: accuracy and reducing execution time. Trying various proportions of GP functions like crossover, mutation, selection, and etc, may result in better accurate GP programs. Applying some statistical methods like Principal Component Analysis(PCA) as preprocessing step and applying some clustering, like Expectation Maximization clustering, may make this approach faster. The further research work will be in the above direction.

References

1. Loveard, T., Ciesielski, V.: Representing classification problems in genetic programming. In: Proc. Congr. Evolutionary Computation, May 27-30, pp. 1070–1077 (2001)
2. Kishore, J.K., Patnaik, L.M., Mani, V., Agrawal, V.K.: Application of genetic programming for multicategory pattern classification. *IEEE Transaction on Evolutionary Computation* 4, 242–258 (2000)
3. Muni, D.P., Pal, N.R., Das, J.: A novel approach for designing classifiers using genetic programming. *IEEE Trans. Evolut. Comput.* 8(2), 183–196 (2004)
4. Muni, D.P., Pal, N.R., Das, J.: Genetic programming for simultaneous feature selection and classifier design. *Systems, Man, and Cybernetics*, *IEEE Transactions on Systems, Man, and Cybernetics, Part B* 36(1), 106–117 (2006)
5. Paul, T.K., Iba, H.: Prediction of Cancer class with Majority Voting Genetic Programming Classifier Using Gene Expression Data. *2009 IEEE/ACM Trans. on Computational Biology and Bioinformatics* 6(2), 363–367 (2009)
6. Han, J., Kamber, M.: *Data Mining Concepts and Techniques*, 2nd edn. Elsevier (2006)
7. Goldberg, D.E.: *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, Reading (1989)
8. Koza, J.R.: *Genetic Programming: On the programming of Computers by Means of Natural Selection*. M.I.T. Press, Cambridge (1992)
9. Poli, R., Langdon, W.B., McPhee, N.F.: *A field guide to Genetic Programming* (March 2008), <http://www.gp-field-guide.org.uk>
10. Duda, R.O., Hart, P.E., Stork, D.G.: *Pattern Classification*, 2nd edn. John Wiley and Sons (2001)
11. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representation by error propagation. In: Rumelhart, D.E., McClelland, J.L. (eds.) *Parallel Distributed Processing*. MIT Press (1986)
12. Breiman, L.: Bagging predictors. *Machine Learning* 24, 123–140 (1996)
13. Tan, P.-N., Steinbach, M., Kumar, V.: *Introduction to Data Mining*. Person Education (2006)
14. Nagendra Kumar, D.J., Satapathy, S.C., Murthy, J.V.R.: A scalable genetic programming multi-class ensemble classifier. In: *World Congress on Nature & Biologically Inspired Computing, NaBIC 2009*, pp. 1201–1206 (2009), doi:10.1109/NABIC.2009.5393788
15. Masud, M.M., Gao, J., Khan, L., Han, J., Thuraisingham, B.: Integrating Novel Class Detection with Classification for Concept-Drifting Data Streams. In: Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (eds.) *ECML PKDD 2009*. LNCS, vol. 5782, pp. 79–94. Springer, Heidelberg (2009)
16. Folino, G., Pizzuti, C., Spezzano, G.: An Adaptive Distributed Ensemble Approach to Mine Concept-Drifting Data Streams. In: *ICTAI 2007 Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence*, vol. 02 (2007)
17. Abbass, H.A., Bacardit, J., Butz, M.V., Llorà, X.: Online Adaptation in Learning Classifier Systems: Stream Data Mining (2004)
18. Zhang, Y., Jin, X.: An automatic construction and organization strategy for ensemble learning on data streams. *ACM SIGMOD Record Homepage archive* 35(3) (September 2006)
19. Wu, W., Gruenwald, L.: Research issues in mining multiple data streams. In: *Stream KDD 2010 Proceedings of the First International Workshop on Novel Data Stream Pattern Mining Techniques* (2010)

20. Bifet, A., Holmes, G., Pfahringer, B., Kirkby, R., Gavaldà, R.: New ensemble methods for evolving data streams. In: 15th ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining (KDD 2009), Paris, France (June 2009)
21. Folino, G., Pizzuti, C., Spezzano, G.: Boosting Technique for Combining Cellular GP Classifiers. In: Keijzer, M., O'Reilly, U.-M., Lucas, S., Costa, E., Soule, T. (eds.) EuroGP 2004. LNCS, vol. 3003, pp. 47–56. Springer, Heidelberg (2004)
22. Paris, G., Robilliard, D., Fonlupt, C.: Genetic Programming with Boosting for Ambiguities in Regression Problems. In: Ryan, C., Soule, T., Keijzer, M., Tsang, E.P.K., Poli, R., Costa, E. (eds.) EuroGP 2003. LNCS, vol. 2610, Springer, Heidelberg (2003)
23. Teredesai, A., Govindaraju, V.: Issues in Evolving GP based Classifiers for a Pattern Recognition Task. In: Proceedings of the 2004 IEEE Congress on Evolutionary Computation, pp. 509–515. IEEE Press (2004)

Intelligent Chaos Controller

A Computational Intelligence Based Approach for Data-Driven Real-World Systems

Jallu Krishnaiah¹, C.S. Kumar², and M.A. Faruqi³

¹ Formerly Research Scholar at IIT Kharagpur
Present address: of R&D, BHEL, Trichy, India
j.krishnaiah@gmail.com

² Robotics and Intelligent Systems Lab, IIT Kharagpur
kumar@mech.iitkgp.ernet.in

³ Formerly Professor at IIT Kharagpur
Present address: Azad Institute of Engineering and Technology Lucknow
aslamfaruqi@gmail.com

Abstract. Recent developments have shown the possibility of constructing Bifurcation Diagrams for real-world chaotic system based on observed data. In the present work we demonstrate how the control can be achieved on the data-driven process/system based on the bifurcation diagram construction capability. In reality many physical and non-physical systems are very difficult to represent using a mathematical form; even if mathematical models exist, it would be a difficult task to build a controller which works in real-time. Moreover, if the considered system is chaotic in nature there are very few methods for are available controlling. On contrary there are large number of Chaos Control techniques when the considered system is/has a mathematical model. Based on the fundamental idea of these techniques, *i.e.* small perturbation at appropriate time is enough to control such a chaotic systems, the present method uses the global search capability of genetic algorithms to find a best perturbation to the control parameter at each step with a RNN model of the considered system as an objective function.

1 Introduction

The recognition that the occurrence of chaos is wide spread in nature, and it even forms the basis of intelligence in brain and provides control mechanism for heart and other systems, and also its prevalence in engineering and scientific systems, has been fueling research in this area. Many books [18,21,7] have been written focusing on different aspects of the chaotic systems. A recent literature review has been provided by Andrieveskii and Fradkov [16] showing the diversity in chaos related research. Further, recent works [29,9] have demonstrated the controllability through chaos concepts.

Most of the chaos controlling methods work on the fact that the relationship at fixed point on a Poincaré section is linear [20]. In these methods instead of a

deliberate action to control the chaos in the system, wait till the system reaches the vicinity of the fixed point. Then the controlling action is applied in the form of a small one-time or continuous perturbation to the accessible control parameter. In reality waiting for the system to reach the vicinity of the fixed point is an indefinite time-taking process. In some cases of chaos control very heavy computing may be necessary to find the magnitude of the perturbation and the step at which the perturbation is required to apply. Moreover, knowing the exact mathematical models for real-world systems is also a difficult proposition, and also they are prone to noise and errors.

Studies of real-world systems have to additionally contend with the prevalence of noise in systems, its role in changing the initial conditions of the system and its capability to cause bifurcations [28]. The problem gets further complicated with time scale of the phenomena to be controlled, if the computations for control actions have to be done in real-time. The three main methods of controlling chaos that have been established are [17]

1. Open loop method based as periodic excitation of the system,
2. OGY method(Poincaré map linearization), and
3. Method of time delayed feedback (Pyragas method).

The open loop method has been in use for a long time specially in systems with high frequency operations, to alter their behaviour with small changes in inputs. Well developed procedures are available today.

It is the OGY method which made the control of chaos possible in large variety of continuous and discrete systems. The method aims at stabilizing the periodic orbits of a system at an unstable fixed point. Today numerous examples [32,8,5] are available of its successful application even in presence of noise [5,25]. The original method requires waiting for suitable orbits to occur before stabilization is attempted.

Pyragas' [26] method has been used successfully for stabilizing chaotic systems without a complete model of the system, through suitably derived time delayed feedbacks to the system [24]. However, formal analysis for understanding the process and the conditions of stability are still ongoing research activities.

The basic problem in identifying the presence of chaos in real-world system has been addressed through identification procedure using observed data [22,10]. Taking into account the literature survey in the sections on the architecture and capability of RNN models, the role of Bifurcation in chaos, and practical control strategies for real-world systems, the following areas have been chosen for further investigations as they have not been sufficiently explored.

2 Review on Techniques

2.1 Recurrent Neural Network

After the discovery of Backpropagation algorithms for Feed-forward neural networks, many researchers explored the possibility of extending the Backpropagation algorithms to the cases of networks with recurrent connections. Pineda [23],

Almeida [2], and Rohwer and Forrest [27] have independently pointed out that Backpropagation can be extended to arbitrary networks as long as they converge to stable states.

Elman in [12], suggested a Recurrent Neural Networks architecture with input layer is divided into two parts: the true input units and the context input units. The context units simply hold a copy of the activations of the hidden layer from the previous time step. The modifiable connections are all feed-forward, and can be trained by conventional Backpropagation methods.

Bengio & F.Gingras [3] proposed recurrent neural networks with feedback into the input units for handling the problems like missing input variables or input variables being available at different frequencies to minimize the output error.

An experimental comparison of some of the variants of recurrent neural networks has been done by Horne and Giles [4] to model a set of Dynamical Systems Equations. The report concludes that in general all the recurrent neural networks perform better when compared to Time Delay Neural Networks (TDNN). Among the RNN models, Narendra and Parthasarathi's model performs better than others.

Bengio et. al. [33] proved in their paper that "Long-term dependencies with gradients is difficult" if the system is to latch on information robustly, then the fraction of the gradient in a gradient-based training algorithm due to information in n time steps in past approaches zero as n becomes large. This effect is called the problem of vanishing gradient. Bengio *et. al.* [33] claimed that the problem of vanishing gradients is essential reason why gradient-descent methods are not sufficiently powerful for long-term dependencies. Tsungnan Lin *et. al.* [30] have recently shown one class of recurrent neural networks called NARX (Nonlinear Auto-Regressive with eXogenous inputs) which has the capability of learning long-term dependencies i.e. those problems for which the desired output of the system at time T depends on inputs present at time $t \ll T$. Here inputs also have time delayed information, in addition to outputs being fed-back. It has been shown that it is also a good candidate for the dynamical systems modelling.

Recently, there has been an interest in understanding the dynamical properties of RNN in terms of connection weight between the nodes. In these studies it has been found that even a two node fully connected RNN model could show a Bifurcation behaviour with respect to connection weights [11,14] *etc.*

2.2 Chaos Control

There are several approaches evolved after the introduction of famous OGY method of chaos control. Till the introduction of OGY method it was assumed that the chaos can not be controlled and one has to avoid the region of chaos and take safer route and keep in non chaotic region. After understating the nature of chaos, it has been shown that the chaos is a co-existence state of many periodicities which are unstable in nature. In the OGY method utilizing the richness of the very property has been demonstrated by controlling to a unstable periodic orbit. That means it is possible to switch to any periodic orbit once the system is in chaotic mode. This gives a wide opportunity of easily changing

the mode with less amount of energy. And also, sometime operating the system in non-chaotic region would be an expensive when comparing to chaotic region. Hence, controlling and stabilizing the system in chaotic region is very important.

2.3 Application of RNN to Chaos Control

While RNN system have been extensively used for modelling complex non-linear dynamical systems their evolution towards representing chaotic systems has been reported only in few publications.

J.C.Principe [16] proposed a method to model a chaotic system and shown it is to predict time-series for a larger number of steps for a given initial condition. The main limitation of this model is that it considers only one set of accessible control parameters, which may not be enough to model and control real-world systems. Another variation of the above approach has been used by Jones *et al.* [19] to model more complex systems for control by providing additional information at feedback input points. Jones *et al.* [13] have explored the possibility of synchronising chaotic systems, using recurrent neural networks, trained to low values of mean squared error of training. The possibilities of using these for real-world system were not explored enough.

From the literature it can thus be seen that recurrent neural networks have been used extensively with various architectures for one step or several steps prediction from a given starting point.

2.4 Genetic Algorithms

Genetic algorithms were formally introduced in the United States in the 1970s by John Holland at University of Michigan. The continuous reduction of price/performance of computational systems has made them attractive for some types of optimization. They are less susceptible to getting 'stuck' at local optima than gradient search methods. But they tend to be computationally expensive in some cases where objective function is heavy. As by now genetic algorithms are common, we are not treating this section in detail.

There has been some work to implement GA to search controller to control chaotic systems. Among these important is Weeks and Burgess [31]. Weeks and Burgess proposed a method for controlling chaos where the controller is a Neural Network model, getting feedback from the plant itself. The Neuro-controller was evolved through a Genetic Algorithm based search. The Neuro-controller gives a perturbation to the control parameters based on last three time states of the system. This method was also basically devoted to control of chaotic system through orbit stabilization procedures [20].

3 Intelligent Chaos Controller

The present framework is inspired by the OGY method. The creative thought that has been implemented in the OGY method and Weeks GA based neuro-controller are the origin to present way of controlling chaos through a RNN

model and GA based approach. In OGY method one has to wait to apply the control action till the system reach the vicinity of the fixed point or the periodic orbit that being sought to achieve i.e. period one, period two period four etc. In the present approach the control action starts immediately.

The present approach is targeted to control real-world chaotic systems. It is known that the real-world systems are difficult to represent through any mathematical equations. They can be represented through the observed multivariate data. A newly developed methodology that could be verified the trueness of the developed model was introduced in [15]. In that, it has been shown that the developed model could be used for constructing the bifurcation diagrams of the system for various input conditions. Further, it has been shown that the developed model could be used for calculating maximum Lyapunov Exponents of system for any given initial conditions. This possibility gave the strength to use those developed models to find a step-by-step controller to stabilize the given system.

The present approach has been tried with two kinds of perturbations. One is uniform perturbation and second is differential (adaptive) perturbation. First one is simple and one range of space for perturbation. Second one works on the multiple ranges of space of perturbation. And the perturbation range is decided based on the fitness of the solutions and the present state of the system.

In general most of the studies in Chaos Theory and related are first experimented on well known discrete systems also known as maps. The present framework is experimented on two such maps namely Hénon and Ikeda maps.

3.1 How to Build RNN-GA Framework?

The main targeted application of the present framework is to build a intelligent Chaos Controller for Industrial Processes/Systems. Therefore the procedure has been developed keeping in mind that how the these processes operated and observed for the data collection. Moreover they are prone to noise and available short (short because some process may take 1hr to 1day for a single set of observation, therefore it may take lot of time to have large set of data).

Once the data for a given process is available, using the generic architecture of Recurrent Neural Networks(RNN) described in [15], model is built. After analyzing the model for its Bifurcation Diagram which represents the behaviour of the given system in-terms of time and as well as control parameters.

Using a standard Genetic Algorithms(GA) for searching the perturbation required to be given the present set of control parameters of the considered system is build. The RNN model in hand will be used as objective function in the GA search.

3.2 Control of Hénon Map

Hénon map is well known discrete system, which represents a two dimensional chaotic system. Most of the ideas in the chaos theory have been first verified and evolved for general chaotic systems. The Hénon map can be represented

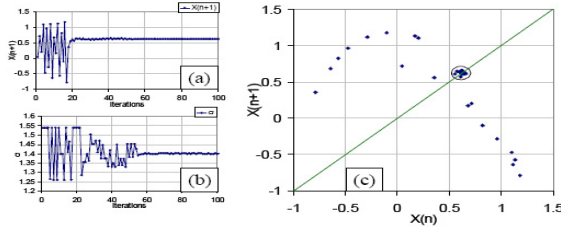


Fig. 1. Controlling chaos in the Hénon map using RNN-GA based approach

by two equations with two constants called control parameters or bifurcation parameters. The controlling a chaotic state of Hénon map through the developed frameworks is shown in Figure.1

4 Controlling Ikeda Map

Ikeda map is another well known discrete chaotic system of dimension three. This map is more complex in its attractor nature. Using the developed RNN-GA based approach the Ikeda map has been tried to stabilize on to a fixed point, though the fixed point is not known. Given some initial conditions, GA suggests amount of perturbation to be given so that the objective function is minimized. In the present test case it took less than 100 iterations to minimize the variation. For the complete stabilization the GA took more that 300 iteration. The controlling a chaotic state of Ikeda map through the developed frameworks is shown in Figure.2

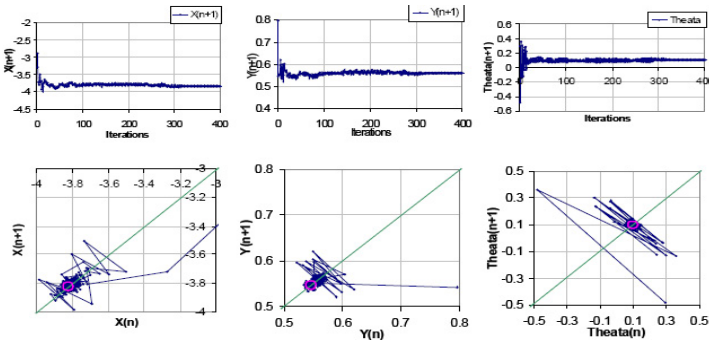


Fig. 2. Controlling chaos in the Ikeda map using RNN-GA based approach

5 Conclusion

The experimentation with these two maps gave encouraging results for the chaos control approach through RNN-GA. It has been observed its simplicity and fastness in stabilizing the system within the chaotic regions, without actually knowing the fixed point. It has been generalized for data driven systems using Recurrent Neural Networks models.

Further, this shows the general applicability of the framework described in [15]. This framework may be applied to any process or system which has complex nonlinear dynamic behaviour.

Acknowledgments. The authors are thankful to the SERC-DST, Govt of India for providing grants as per Sanction no:SR/FTP/ET-83/2001 under the scheme of Fast-track proposal for young scientists.

References

1. Fradkov, A.L., Evans, R.J.: Control of chaos: Survey 1997-2000. In: Preprints of 15th IFAC World Congress on Automatic Control. Plenary papers, Survey papers, Milestones, Barcelona, pp. 143–154 (July 2002)
2. Almeida, L.: Backpropagation in perceptrons with feedback. In: Eckmiller, R., der Malsburg, V. (eds.) *Neural Computers*, pp. 199–208. Springer, New York (1988)
3. Bengio, Y., Gingras, F.: Recurrent neural networks for missing or asynchronous data. In: *Neural Information Processing Systems*. MIT Press (1996)
4. Horne, B.G., Giles, C.: An experimental comparison of recurrent neural networks. In: *Advances in Neural Information Processing Systems*, p. 697. MIT Press (1995)
5. Boccaletti, S., Grebogi, C., Lai, Y.C., Mancini, H., Maza, D.: The control of chaos: Theory and applications. *Physics Reports* 329(3), 103–197 (2000)
6. Andrievskii, B.R., Fradkov, A.L.: Control of chaos: Methods and applications. 2. applications. *Automation and Remote Control* 65(4), 505–533 (2003)
7. Hilborn, R.C.: *Chaos and Nonlinear Dynamics: An Introduction for Scientists and Engineers*, 2nd edn. Oxford University Press, New York (2000)
8. Christini, D.J., Collins, J.J.: Controlling neuronal chaos using chaos control, <http://arxiv.org/abs/chao-dyn/9503003>
9. Dattani, J., Blake, J.C., Hilker, F.M.: Target-oriented chaos control. *Physics Letters A* 375(45), 3986–3992 (2011), <http://www.sciencedirect.com/science/article/pii/S0375960111011194>
10. De Feo, O.: Self-emergence of chaos in the identification of irregular periodic behaviour. *CHAOS* 13(4), 1205–1215 (2003)
11. Doya, K.: Bifurcations in the learning of recurrent neural networks. In: *Proc. IEEE Int. Symp. Circuits and Systems*, pp. 2777–2780. IEEE (1992)
12. Elman, J.: Finding structure in time. *Cognitive Science* 6(2), 285–324 (1990)
13. Jones, A.J., Tsui, A., Oliveira, A.G.: Neural models of arbitrary chaotic systems: construction and the role of time delayed feedback in control and synchronization. *Complexity International* 9 (2002)
14. Aihara, K., Takabe, T., Toyoda, M.: Chaotic neural networks. *Special Issue on Recurrent Networks for Sequence Processing* 144(12), 333–340 (1990)

15. Krishnaiah, J., Kumar, C., Faruqi, M.A.: Modelling and control of chaotic processes through their bifurcation diagrams generated with the help of recurrent neural networks models: Part 1 - simulation studies. *Journal of Process Control* 16(1), 53–66 (2006)
16. Kuo, J.M., Principe, J.C., de Vries, B.: Prediction of chaotic time-series using recurrent neural networks. Submitted to IEEE Workshop NN for SP (1992)
17. Fradkov, A.L., Evans, R.J.: Control of chaos: Survey 1997–2000. Survey paper. Univ. of Melbourne, St. Petersburg (2002)
18. Nayfeh, A.H., Balachandran, B.: *Applied Nonlinear Dynamics: Analytical, Computational and Experimental Methods*. John Wiley and Sons, New York (1995)
19. de Oliveira, A.G., Tsui, A.P., Jones, A.J.: Using a neural network to calculate the sensitivity vectors in synchronisation of chaotic maps. In: *Proceedings 1997 International Symposium on Nonlinear Theory and its Applications (NOLTA 1997)*, vol. 1, pp. 46–49. Research Society of Nonlinear Theory and its Applications, IE-ICE, Honolulu, U.S.A (1997)
20. Ott, E., Grebogi, C., Yorke, J.A.: Controlling of chaos. *Phys. Rev. Lett.* 64, 1192–1196 (1990)
21. Ott, E.: *Chaos in Dynamical Systems*. Cambridge University Press, Maryland (1993)
22. Panzyak, A., Yu, W., Sanchez, E.: Identification and control of unknown chaotic systems via dynamical neural networks. *IEEE Trans. on Circuits and Systems* 46(12), 1491–1495 (1999)
23. Pineda, F.J.: Generalization of backpropagation to recurrent neural networks. *Physical Review Letters* 59, 2229–2232 (1987)
24. Tsui, A.P.M., Jones, A.J.: Periodic response to external stimulation of a chaotic neural network with delayed feedback. *Int. J. of Bifurcation and Chaos* 9(4), 713–722 (1999)
25. Po-Feng, Chu, J.Z., Jang, S.S., Shieh, S.S.: Developing a robust model predictive control architecture through regional knowledge analysis of artificial neural networks. *Journal of Process Control* 13, 423–435 (2002)
26. Pyragas, K.: Continuous control of chaos by self-controlling feedback. *Physics Letters A* 170, 421–428 (1992)
27. Rohwer, R., Forrest, B.: Training time-dependence in neural networks. In: *Proceedings of the First IEEE International Conference on Neural Networks*, San Diego, CA, vol. 2, pp. 701–708 (1987)
28. Schenk-Hoppé, K.R.: Bifurcations of the randomly perturbed logistic map - numerical study and visualizations,
<http://www.iew.unizh.ch/home/klaus/logistic/intro.html>
29. Schöll, E.: Neural control: Chaos control sets the pace. *Nature Physics* (2010)
30. Lin, T., Horne, B.G., Tino, P., Giles, C.: Learning long-term dependencies in narx recurrent neural networks. *IEEE Trans. on Neural Networks* 7(6), 1329 (1996)
31. Weeks, E.R., Burgess, J.M.: Evolving artificial neural networks to control chaotic systems. *Physical Review E* 56(2), 1531–1540 (1997)
32. Ditto, W.L., Rauseo, S.N., Spano, M.L.: Experimental control of chaos. *Phys. Rev. Lett.* 65(26), 3211–3214 (1990)
33. Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient is difficult. *IEEE Transactions on Neural Networks* 5(2), 157–166 (1994)

Protein Structure Prediction in 2D HP Lattice Model Using Differential Evolutionary Algorithm

Nanda Dulal Jana¹ and Jaya Sil²

¹Department of Information Technology,
National Institute of Technology, Durgapur,
West Bengal, India

²Department of Computer Science and Technology,
Bengal Engineering & Science University, Shibpur, West Bengal, India
nanda.jana@gmail.com, js@cs.becs.ac.in

Abstract. Protein Structure Prediction (PSP) is a challenging problem in bioinformatics and computational biology research for its immense scope of application in drug design, disease prediction, name a few. Developing a suitable optimization technique for predicting the structure of proteins has been addressed in the paper, using Differential Evolutionary (DE) algorithm applied in the square 2D HP lattice model. In the work, we concentrate on handling infeasible solutions and modify control parameters like population size (NP), scale factor (F), crossover ratio (CR) and mutation strategy of the DE algorithm to improve its performance in PSP problem. The proposed method is compared with the existing methods using benchmark sequence of protein databases, showing very promising and effective performance in PSP problem.

1 Introduction

One of the greatest challenges in bioinformatics research is to solve protein folding problem, called protein structure prediction from its primary amino acids sequences. A protein is represented by a sequence of 20 different amino acids, joined end to end by formation of peptide bonds. Fig. 1 shows the peptide bond between two amino acids.

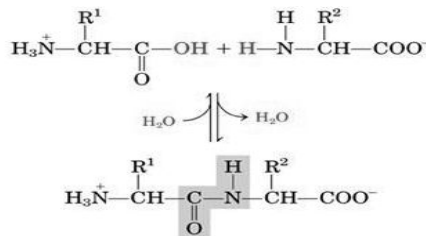


Fig. 1. Peptide Bonds among two amino acids

The 3D structure (native structure) of a protein describes biological functions, which play an important role in drug design, disease prediction and so on. Biological scientists predict the structure of proteins by experiments like X-ray crystallography and nuclear Magnetic Resonance (NMR) [1]. However, these processes are very time consuming and expensive too, so researchers concentrate on protein structure prediction using computational strategies.

Even for a small protein sequence, exhaustive search is impossible due to the exponential growth in the number of possible conformations with the number of amino acids. Moreover, the computational analysis of the prediction of structures is intractable by using simple lattice models [2]. To overcome the limitations, researchers use a heuristic optimization method, in particular evolutionary algorithms [3, 4, 5] to predict 3D protein structure. In this work, simple 2D HP lattice model [6] has been considered where amino acids are characterized by polar (P) and non-polar (H) residues of amino acid. In this model, each H and P is embedded on 2-D square lattice with non-overlapping amino acids, called feasible conformation. In infeasible conformation, amino acids are overlapping on lattice. The total numbers of hydrophobic contacts i.e., H-H non local contacts between the amino acids, which are not adjacent in the sequence are used as energy function in this model.

In the paper, a Differential Evolutionary (DE) algorithm for protein structure prediction (PSP) problem based on the 2D HP lattice model has been presented. First infeasible conformation is converted to feasible conformations by checking possible relative movement of the amino acids. To improve the performance of the DE algorithm, selections of control parameters such as NP, F, and CR are modified. Finally, results produced by this algorithm are compared with previously published results.

The paper is structured as follows: Section 2 presents the preliminaries of 2D HP lattice model and section 3 describe the Differential Evolutionary Algorithm briefly. Methodology for applying DE algorithm to PSP problem is described in section 4. In section 5, the experimental results are compared against other known algorithms. Finally, conclusion and future direction is summarized in section 6.

2 2D HP Model

The most widely used discrete model for protein structure prediction is 2D HP lattice model [6]. In this model each amino acid is classified as hydrophobic or non-polar (H) or a hydrophilic or polar (P) based on their hydrophobicity. Conformation of a protein is then represented as a self-avoiding walk i.e., a feasible conformation in a 2D HP square lattice. The basic concept of this model is that the hydrophobic (H) amino acids lying in its core to provide more stable structure with minimum free energy. Each hydrophobic (H) amino acids tend to avoid interact with solvent environment and hence tend to move inside the structure where polar amino acids remain on the outside of the structure.

An H-H non local bond is a pair of Hs that are adjacent in the lattice but not in the sequence. The native conformation of a protein corresponds to the minimum free energy conformation for that protein. The optimal feasible conformation in the square 2D HP lattice model is one that has the maximum number of H-H non local bonds,

which give minimum energy value. In Fig. 2, the black circles and the white circles represent hydrophobic (H) amino acids and hydrophilic (P) amino acids respectively. ‘S’ and ‘E’ represent the starting amino acid and end amino acid while dotted lines represent H-H non local contacts. The conformation has 9 H-H non local contacts and this is the maximum number of contacts for the given sequence.

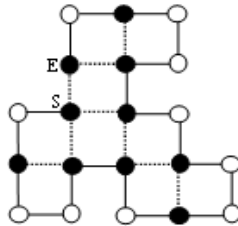


Fig. 2. An optimal conformation for the sequence HPHPPHHPHPPHPPHPPH in the 2D square lattice

3 Differential Evolutionary (DE) Algorithm

Differential Evolutionary (DE) algorithm [7, 8, 9] is a population based powerful stochastic search method for global optimization problem. It is applied on the optimization problem when the possible solutions are represented by a real-valued vector. In this algorithm, an individual is a set of D optimization parameters, represented by a D-dimensional parameter vector. Like other evolutionary algorithms, DE represented by an initial population (P), is a set of NP, D-dimensional parameter vectors. The generation G is denoted by G=0, 1, 2, ..., G_{max}. The ith vector of the population in the current generation G is given as-

$$\bar{X}_{i,G} = \{x_{1,i,G}, x_{2,i,G}, x_{3,i,G}, \dots, x_{D,i,G}\}$$

Here $x_{j,i,G}$ is the jth component (j=1 ... D) of the ith parameter vector (i=1 ... NP) at generation G and the value of the parameter is randomly generated using a uniform distribution $x_j^{low} \leq x_{j,i,G} \leq x_j^{up}$, where x_j^{low} and x_j^{up} is the lower and upper limit of $x_{j,i,G}$. After initialization of the parameter vectors, DE enters in a cycle to execute the steps: mutation, crossover and selection as described below.

3.1 Mutation

Mutation operation creates a donor vector $\bar{V}_{i,G}$ corresponding to each population or target vector $\bar{X}_{i,G}$ in the current generation G. Different mutation strategies used in several optimization problems are described in [10]:

DE/rand/1:
$$\bar{V}_{i,G} = \bar{X}_{r1,G} + F \cdot (\bar{X}_{r2,G} - \bar{X}_{r3,G})$$

DE/best/1:
$$\bar{V}_{i,G} = \bar{X}_{best,G} + F \cdot (\bar{X}_{r1,G} - \bar{X}_{r2,G})$$

$$\begin{aligned} \text{DE/best/2:} \quad & \bar{v}_{i,G} = \bar{x}_{best,G} + F \cdot (\bar{x}_{r1,G} - \bar{x}_{r2,G}) + F \cdot (\bar{x}_{r3,G} - \bar{x}_{r4}) \\ \text{De/rand/2:} \quad & \bar{v}_{i,G} = \bar{x}_{r1,G} + F \cdot (\bar{x}_{r2,G} - \bar{x}_{r3,G}) + F \cdot (\bar{x}_{r4,G} - \bar{x}_{r5}) \end{aligned}$$

Where the indices $r_1^i, r_2^i, r_3^i, r_4^i, r_5^i$ are mutually exclusive integers, randomly chosen from the range [1, NP] and different from the base vector i . For each donor vector, these indices are generated randomly. Here $F \in [0, 2]$ [11] is a positive control parameter, known as scaling factor used to scale the difference vectors. \bar{x}_{best} is the best individual vector in the population at generation G .

3.2 Crossover

Crossover operation plays an important role to explore the search space in DE. The crossover operation is applied to each pair of the target vector $\bar{X}_{i,G}$ and its corresponding donor vector $\bar{V}_{i,G}$ to generate a trial vector $\bar{U}_{i,G} = \{u_{1,i,G}, u_{2,i,G}, u_{3,i,G}, \dots, u_{D,i,G}\}$. In DE, there are two types of crossover techniques: binomial and exponential crossover [7]. In binomial crossover, the trial vector obtained as

$$\bar{U}_{j,i,G} = \begin{cases} v_{j,i,G} & \text{if } (rand \leq CR \text{ or } j = j_{rand}) \\ x_{j,i,G} & \text{otherwise} \end{cases}$$

Where $rand \in [0, 1]$, is a uniformly distributed random number for j^{th} component of i^{th} parameter vector. $CR \in [0, 1]$ is a crossover probability, which defines by user that controls the parameter values are copied from the donor vector. $j_{rand} \in [1, D]$ is a randomly chosen index, which ensures that \bar{U}_j gets at least one component from $\bar{V}_{i,G}$

In exponential crossover, first an integer $n \in [1, D]$ is chosen randomly. This integer acts as a starting point in the target vector from where the crossover or exchange of components starts with the donor vector. Another integer L is chosen from the interval [1, D] where L denotes the number of components the donor vector actually contributes to the target. After choosing n and L , the trial vector is obtained as

$$\bar{U}_{j,i,G} = \begin{cases} v_{j,i,G} & \text{for } j = \langle n \rangle_D, \langle n + 1 \rangle_D, \langle n + 2 \rangle_D, \dots, \langle n + L - 1 \rangle_D \\ x_{j,i,G} & \text{for all other } j \in [1, D] \end{cases}$$

Where $\langle \cdot \rangle_D$ denote modulo function with modulus D . The integer $L \in [1, D]$ is taken according to the following pseudo-code:

```
L=0; Do
{ L=L+1; } While ((rand (0, 1) < CR) and (L < D))
```

Where CR is the crossover probability. Hence in effect, probability $(L \geq v) = (CR)^{v-1}$ for any $v > 0$.

3.3 Selection

The next stage of DE algorithm is selection operation to determine whether the target vector or the trial vector survives to the next generations i.e., at $G=G+1$. The selection operation is describe as

$$\vec{x}_{i,G+1} = \begin{cases} \vec{U}_{i,G} & \text{if } f(\vec{U}_{i,G}) \leq f(\vec{x}_{i,G}) \\ \vec{x}_{i,G} & \text{otherwise} \end{cases}$$

Where $f(X)$ is the function to be minimized. Depending on the stopping criteria, the above three stages are repeated generation after generation.

4 Methodology

4.1 Vector Encoding

The performance of an evolutionary algorithm is strongly depending on the way of representing the individuals, a set of optimization parameters. To date, there are three ways to represent a conformation of a protein on a lattice [6]: Distance matrix, Cartesian coordinates and internal coordinates. In this work, internal coordinates are used, which are two types [12]: Absolute internal coordinate and Relative internal coordinate. In absolute internal coordinate, according to the axis of the lattice, a given amino acid moves. The conformation, using this scheme are coded with a sequence in $\{N, S, E, W\}^{n-1}$, which corresponds to North, South, East and West for n length protein sequence in 2D lattice. But, in Relative internal coordinate encoding, a given amino acid moves according to the previous amino acid movements. Using this encoding scheme, conformation is represented with a sequence in $\{F, L, R\}^{n-1}$, which corresponds to Forward, Left and Right for n length protein sequence in the plain.

In DE, every individual are real valued vectors, decoded into a specific conformation of a protein on the 2D square lattice. Therefore, an adaptation concept is necessary for encoding and decoding the sequence of movements of a protein on the lattice. The same adaption concept proposed in [5] has been used in the paper. Using relative internal coordinate in 2D square HP lattice model, the movements are Forward, Left and Right. Therefore, the phenotypical representation of a solution is defined over the alphabets $\{F, L, R\}$. The genotypical representation is still a real valued vector. Consider $x_{i,j}$ is the j^{th} element of individual and P is the string representing the sequence of movements of the conformation and $\alpha < \beta < \gamma < \delta$ arbitrary constants in \mathbb{R} . The genotype-phenotype mapping is defined as follows:

$$\text{if } \alpha < x_{i,j} \leq \beta \text{ then } P_j = L$$

$$\text{if } \beta < x_{i,j} < \gamma \text{ then } P_j = F$$

$$\text{if } \gamma \leq x_{i,j} < \delta \text{ then } P_j = R$$

In this work, $\alpha = -3, \beta = -1, \gamma = 1, \delta = 3$ are considered as in [5].

4.2 Initial Population

Initial population has been generated randomly using the relative internal coordinate encoding scheme. Therefore, the conformation of a protein is represented by a string of alphabets F, L, and R. Using this scheme, the conformations of proteins may be feasible (non overlapping of amino acids on the lattice) or infeasible (overlapping of amino acids). Thus, using the above defined genotype-phenotype mapping we convert the string of conformation to real valued vector, because in Differential Evolutionary algorithm every individual is real valued vector. When evaluating the fitness (maximum number of H-H non local bonds) of a conformation, again the individual (real valued vector) is converted to string of alphabets F, L, and R using the same genotype-phenotype mapping. We assume, the infeasible conformations are given a fitness of -1 and a mechanism is proposed to convert the infeasible conformation to feasible one.

4.3 Proposed Mechanism

Basically, there is no fixed technique for converting infeasible conformation to feasible ones. In this mechanism, an infeasible protein conformation (string of characters F, L, R) is taken as inputs. First, we check a movement of the string one by one from the starting movement to end of infeasible conformation to check whether conflict (existences of overlapping) is occurred or not. If any conflict occurred with the movement, then we check the possible movement except the current movement resulting nonoccurrence of conflict. If one possible movement exists, we replace the current movement by finding new movement and rest of the movements is unchanged in which no conflict occurs. If there is two possible movements exist, select any one arbitrarily. This checking procedure is repeated through the rest of the movements in the infeasible conformation. If there is no possible movement, we consider this conformation is an infeasible conformation and assign the fitness to -1. Since, our objective is to maximize the number of H-H non local bonds using the DE therefore, after some generation infeasible conformation has been removed from the population. The proposed algorithm is shown in Fig. 3. In Fig. 4, (a) is an infeasible conformation with string 'FFLLRLLFLR'. The movement F creates a conflict with 1st and 9th amino acids. There are two movements: L and R are to be checked. But L movement also creates a conflict with 5th amino acid. Therefore, only R movement is possible as shown in (b) where the feasible conformation is FFLLRLLRLR. In Fig. 4, (c) is the infeasible conformation where (d) and (e) are the two possible feasible conformations. In Fig. 4, (f) is an infeasible conformation with string 'FLFLFLLFF'. The movement F creates a conflict with 3rd and 9th amino acids. Two movements L and R conflict with 5th and 1st amino acid and so this remains as infeasible conformation.

4.4 DE control Parameters

The mutation strategy, crossover strategy and control parameters such as the population size (NP), crossover ratio (CR) and the scale factor (F) are strongly influence the performance [10, 13] of the DE algorithm. Therefore, it is necessary for appropriate combination of strategy and their associated parameter values to solve specific optimization problems. In DE, larger population size explores the

search space but decrease the probability to find the correct search direction. In this work, we used the population size (NP) from 5D to 10D (D is the dimension of the problem) [8].

```

begin
  for i = 1 to n
    Check {current movement} of an individual (string of F, L, R) to
      certain protein conformation produced conflict or not
    if yes then
      Find other possible movement  $S = \{F, L, R\} - \{\text{current movement}\}$ 
      if movement exists then
        {Current movement} is replaced by  $S_1$  where  $S_1 \in S$ .
      else
        break
      Go to next individual
    end
  else
    Go to next individual
  end
end
end

```

Fig. 3. Algorithm for converted infeasible conformation to feasible conformation

The exploration and exploitation of the DE algorithms is very sensitive to the selection of mutation strategy. The donor vectors are created using mutation strategy. The most widely used strategy are DE/rand/1/- and DE/best/1/-. The first strategy is responsible for exploring the search space and the other is used for fast convergence to global optima. Initially, we used DE/best/1/- strategy but if no improvement in best fitness have been seen with N number of generations, then change to strategy DE/rand/1/- up to M number of generations. If fitness is improved within M generations, back to the initial strategy, otherwise back to the initial strategy after M generations. Here, we also consider one difference vector to be perturbed because more difference vectors increase the convergence speed at the cost of possibility to trap at local optima. The scale factor (F) has great importance to the DE algorithm. The large values of F are used for escaping the solution from a local optimum and small values provide rapid convergence but high probability to trap to local optima. Therefore, we used F value from 0.5 to 0.9 at each generation to generate the donor vector. If some components of the donor vector violate its limits, then set the corresponding component to a random value within the specified limits of that component.

In this paper, exponential crossover with crossover probability (CR) from 0.8 to 1. Since, large crossover rate speed up the convergence [11]. Here objective is to find the maximum number of H-H non local contacts.

5 Results and Discussion

In this section, we explain the results obtained by the improved DE algorithm on various benchmark sequence [14] and compare them with the results of protein structure prediction by Genetic algorithm [14], Multimeme Algorithm [15], DE approach [5] and

hybrid DE [16]. We are considering 50 runs for each benchmark sequence using different random seeds. For the experiments, we used the following parameters: $NP \in [5D, 10D]$, $F \in [0.5, 1]$ and $CR \in [0.8, 1]$. To explore the search space, alternatively use the strategy DE/best/1/exp. and DE/rand/1/exp. Using the above strategy adaption, we consider $N=50$ and $M=70$. The algorithm was developed in MatLab 2010b and run on a PC 2.26 GHz core 2 duo with 2 GB RAM under Windows XP.

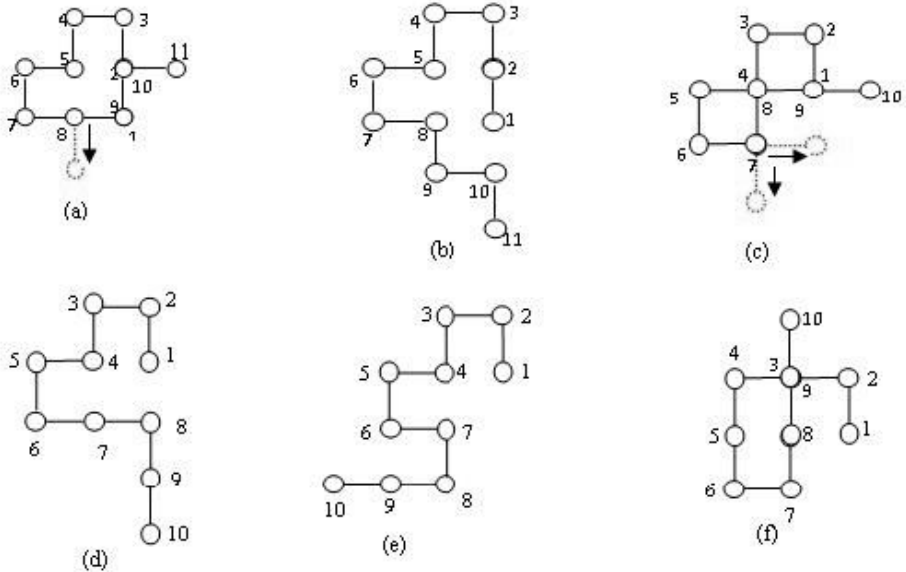


Fig. 4. (a) infeasible conformation of 11 lengths protein sequence; (b) feasible conformation of (a); (c) infeasible conformation with two possible movements; (d) and (e) are the two feasible conformations of (c) and (f) infeasible conformation

The benchmark sequences are shown in Table 1. These sequences of proteins are not the real world proteins but benchmark for 2D HP square lattice model. In Table 1, H^i , P^i and $(HP)^i$ represents the repetitions of the respective amino acids while C_{max} represents the maximum number of H-H non local contacts known to date.

Table 2 shows the results of the proposed approach and other evolutionary algorithmic approaches. In this table, 1st, 2nd and 3rd column shows the sequence number, length of the sequence and maximum (C_{max}) H-H non local contacts respectively. The 4th, 5th, 6th and 7th column represent C_{max} using Genetic Algorithms [14], Multimem Algorithms (MMA) [15], Differential Evolution approach [5] and hybrid DE [16]. Blank space in 5th column represents that the corresponding sequence is not considered. Last column, split by two: first, maximum H-H contacts are obtained and the number of times this maximum was found within the parenthesis in 50 independent runs. Next, the average number over 50 independent runs is listed in the last column.

Result using the proposed approach is better or equal than the GA technique for all the sequences. For the sequence S1, S3, S4, S6 and S8 have equal C_{max} in both MMA and the proposed one. For S5, we obtained better result over MMA while C_{max} are same with the results by hybrid DE and DE approach except for the sequence S8.

Table 1. Benchmark sequence for 2D HP square lattice

Seq.No	HP Chain	Length	C_{\max}
S1	HPHP ² H ² PHP ² HPH ² P ² HPH	20	9
S2	H ² P ² HP ² HP ² HP ² HP ² HP ² HP ² H ²	24	9
S3	P ² HP ² H ² P ⁴ H ² P ⁴ H ² P ⁴ H ²	25	8
S4	P ³ H ² P ² H ² P ⁵ H ⁷ P ² H ² P ⁴ H ² P ² HP ²	36	14
S5	P ² HP ² H ² P ² H ² P ⁵ H ¹⁰ P ⁶ H ² P ² H ² P ² HP ² H ⁵	48	23
S6	H ² PHPHPHPH ⁴ PHP ³ HP ³ HP ⁴ HP ³ HP ³ HPH ⁴ PHPHPHPH ²	50	21
S7	P ² H ³ PH ⁸ P ³ H ¹⁰ PHP ³ H ¹² P ⁴ H ⁶ PH ² PHP	60	36
S8	H ¹² PHPHPH ² H ² P ² H ² P ² HP ² H ² P ² H ² P ² HP ² H ² P ² H ² P ² HPH ² PH ¹²	64	42

Table 2. Comparison of Results using different approaches

Seq. No.	Length	C_{\max}	GA[14]	MMA [15]	Hybrid DE[16]	DE[5]	Our Approach	
							Max	Average
S1	20	9	9	9	9	9	9(50)	9.00
S2	24	9	9		9	9	9(50)	9.00
S3	25	8	8	8	8	8	8(50)	8.00
S4	36	14	14	14	14	14	14(50)	14.00
S5	48	23	22	22	23	23	23(45)	22.88
S6	50	21	21	21	21	21	21(50)	21.00
S7	60	36	34		35	35	35(42)	34.82
S8	64	42	37	39	42	42	39(40)	38.80

In this work, we considered smaller population size (NP), random scale factor (F) and random crossover rate (CR) within the defined range. These are the different from hybrid DE and DE approach in which they considered large population size and fixed F and CR value. Also, we proposed a mechanism which is different from the repair process in hybrid DE that converts the infeasible conformations to feasible conformations. consequently, our algorithms took few seconds to complete one run up to the 50 length sequence and for 60 and 64 took average time 150 to 1000 seconds per run.

6 Conclusion and Future Work

In this paper, we proposed an improved DE algorithm for protein structure prediction using the 2D HP square lattice model. Our algorithm combines with the mechanism that converts infeasible conformation to feasible conformation. Random values of Scale Factor (F) and Crossover Ratio (CR) within the specific limits improves the performance of DE algorithm. Selection of small population size (NP) gives faster run within a specific generation. Experimental results on the benchmark sequences show that the proposed approach is promising and effective than GA and MMA and also

from standard DE approach with respect to NP and number of generations. We would like to improve the performance of DE algorithm using Neighborhood Search concepts for large sequence length of proteins and like to use the DE to predict the structure of a protein on the triangular lattice model.

References

1. Unger, R.: The genetic algorithm approach to protein structure prediction. *Structure and Bonding* 110, 153–175 (2004)
2. Berger, B., Leight, T.: Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete. *Journal of Computational Biology* 5(1), 27–40 (1998)
3. Unger, R., Moulton, J.: A Genetic Algorithm for Three Dimensional Protein Folding Simulations. In: *Proceedings of the 5th Annual International Conference on Genetic Algorithms*, pp. 581–588 (1993)
4. Pedersen, J.T., Moulton, J.: Protein Folding Simulations with Genetic Algorithms and a Detailed Molecular Description. *J. Mol. Biol.* 269, 240–259 (1997)
5. Bitello, R., Lopes, H.S.: A differential evolution approach for protein folding. In: *Proc. IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pp. 1–5 (2006)
6. Dill, K.A.: Theory for the folding and stability of globular proteins. *Biochemistry* 24, 1501 (1985)
7. Storn, R.M., Price, K.V.: Differential Evolution- a simple and efficient adaptive scheme for global optimization over continuous spaces, Technical Report TR-95-012, International Computer Science Institute, Berkeley, USA (1995)
8. Storn, R.M., Price, K.V.: Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization* 11(4), 341–359 (1997)
9. Storn, R.M., Price, K.V., Lampinen, J.A.: *Differential Evolution – A Practical Approach to Global Optimization*. Springer, Berlin (2005)
10. Das, S., Suganthan, P.N.: Differential Evolution: A survey of the state-of-the-art. *IEEE Transaction on Evolutionary Computation* 15(1) (2011)
11. Storn, R.: On the usage of differential evolution for function optimization. In: *Biennial Conference of the North American Fuzzy Information Processing Society (NAFIPS)*, pp. 519–524. IEEE, Berkeley (1996)
12. Krasnogor, N., Hart, W.E., Smith, J., Pelta, D.A.: Protein structure prediction with evolutionary algorithms. In: *Proc. Int. Genetic and Evolutionary Computation Conf.*, pp. 1596–1601 (1999)
13. Liu, J., Lampinen, J.: On setting the control parameter of the differential method. In: *Pro. 8th Int., Conf. Soft Computing (MENDEL 2002)*, pp. 11–18 (2002)
14. Unger, R., Moulton, J.: Genetic Algorithms for protein folding simulations. *Journal of Molecular Biology* 231(1), 75–81 (1993)
15. Krasnogor, N., Blackburne, B.P., Burke, E.K., Hirst, J.D.: Multimeme Algorithms for Protein Structure Prediction. In: Guervós, J.J.M., Adamidis, P.A., Beyer, H.-G., Fernández-Villacañás, J.-L., Schwefel, H.-P. (eds.) *PPSN 2002. LNCS*, vol. 2439, pp. 769–778. Springer, Heidelberg (2002)
16. Santos, J., Diéguez, M.: Differential Evolution for Protein Structure Prediction Using the HP Model. In: Ferrández, J.M., Álvarez Sánchez, J.R., de la Paz, F., Toledo, F.J. (eds.) *IWINAC 2011, Part I. LNCS*, vol. 6686, pp. 323–333. Springer, Heidelberg (2011)

Neighborhood Search Operator Tuned Differential Evolution for Solving Non Convex Economic Dispatch Problem

J. Jasper and T. Aruldoss Albert Victoire

Department of Electrical Engineering
Anna University of Technology, Coimbatore

Abstract. This article addresses a novel and effective algorithm for solving the economic load dispatch (ELD) problem of generating units. Generator constraints, such as valve point loading, ramp rate limits and prohibited operating zones are taken into account in the problem formulation of ELD. The cost function of the generating units exhibits nonconvex characteristics, as valve-point effects are modeled and included as rectified sinusoid components in its conventional formulation. The paper investigates the application of neighborhood search operator (NSO) to tune Differential Evolution (DE) algorithm for solving ELD problem considering non-smooth characteristics (NSELD). The objective of the presented method is to perform a neighborhood search for each population member and to accelerate towards finding the global solution. The method is also allowed to explore the search space for new promising areas by replacing weak solutions with randomly selected individuals. The idea of neighborhood search increases the exploitation ability, whereas the replacement feature improves the exploration ability of the technique. To demonstrate its efficiency and feasibility, the NSO tuned DE is applied to solve NSELD problem of power systems with 6 and 13 units. The simulation results obtained from the NSO tuned DE was compared to those from previous literature in terms of solution quality and computational efficiency. It is shown that, the proposed technique for non-convex ELD problem generates quality solutions reliably.

Keywords: Differential evolution, neighborhood search operator, replacement, non-convex economic dispatch problem.

1 Introduction

Non-smooth economic load dispatch (NSELD) problem is one of the important optimization problems which aim at real time allocation of power demand among the online generating units, assuming that a thermal unit commitment has been previously done. The objective of NSELD is to allocate the total generation required to match load at minimal production cost, satisfying all physical and operational constraints [1]. The majority of the existing literature addresses approximated economic load dispatch (AELD) with smooth cost function. Several deterministic methods, such as

lambda iteration method, linear programming, base point participation factor, gradient method and Newton method can effectively solve the above AELD problem [2], [3]. These methods require an approximation that incremental cost curves are monotonically increasing in nature.

On the contrary, NSELD procedure can cope with discontinuities and high order nonlinearities due to ramp rate limits, valve point loading and prohibited operating zones. This makes NSELD problem a non-convex one with nonmonotonically increasing incremental cost function [5]-[8]. Thus, the NSELD has been recognized as not only a more accurate formulation of the ELD problem but also a difficult multimodal problem. The practical NSELD problem is represented as a non-convex optimization problem, which cannot be directly solved by mathematical methods. Over the past decade, many salient methods have been developed to solve these problems, such as evolutionary programming (EP) [5], [7], genetic algorithm (GA) [8], particle swarm optimization (PSO) [9], Tabu search (TS), simulated annealing (SA) [8], ant colony optimization (ACO) and differential evolution (DE) [10].

Probabilistic methods do not always guarantee discovering the global optimum solution in finite time since updating their candidate's position in the solution space requires probabilistic rules. Therefore, fine tuning of the above techniques were applied for every improvement in the solution. Plenty of literature work has been done on the fine tuning of above methods. To be concise, few are included in the reference [4], [5], [8]. Hybrid optimization algorithms are the recent contributions for solving the NSELD problem, which are the combinations of heuristic and deterministic techniques. Hybridization of EP with SQP (EP-SQP) [9], PSO with SQP (PSO-SQP) [9], chaotic differential evolution with SQP (DEC-SQP) [11], and GA with DE [12] are few examples.

In DE, the fittest of an offspring competes one by one with that of the corresponding parent, which is different from the other evolutionary algorithms (EAs). This one by one competition gives rise to a faster convergence rate. However, this faster convergence also leads to a higher probability of obtaining a local optimum because the diversity of the population descends faster during the solution process. To overcome this, a hybrid method is presented, such that a fine tuning of DE is employed. This method explores a better valley of solution during the progress of each run which improves the performance of the probabilistic method, and thus, the possibility of exploring the global solution is increased.

In this article a neighborhood search operator (NSO) is used to fine tune the DE for improvement in quality of the solution. This NSO enhances the performance of the hybrid method in the complex solution space. In addition to this, the hybrid technique abandons the individual with poor fitness value, which is replaced by a randomly selected new individual. This improves the exploring ability of the hybrid method and thereby a better valley of solution can be effectively explored. Two types of NSELD problems namely ELDVL (ELD with valve point loading) and ELDRPZ (ELD with ramp rate limits and prohibited operating zones) of 6-unit system and 13-unit system with fuel cost functions taking into account the valve-point loading effect is tested to validate the feasibility and effectiveness of the proposed method.

2 Problem Formulation

A practical non-smooth type of ELD problem is modelled in this paper. The practical NSELD problem considers nonlinearities such as valve point effect, prohibited operating zones and ramp rate limits. This paper addresses two different nonconvex ELD formulations, which reflect the real-time operating conditions.

2.1 Objective Function

The objective function of the ELD is written as

$$F_T = \min \left(\sum_{i=1}^N F_i(P_i) \right) (\$)$$

Where,

$$F_i(P_i) = \min \left(\sum a_i P_i^2 + b_i P_i + c_i \right) (\$/hr), i = 1, 2, 3, \dots, N \quad (1)$$

where a_i , b_i and c_i are fuel-cost coefficients of the generating unit i , P_i is the power output of the generating unit i in megawatts and F_T is the total fuel cost of all generating units.

Subject to the following constraints

Real power balance

$$\sum_{i=1}^N P_i - P_D - P_L = 0 \quad (2)$$

where P_D is the power demand and P_L is the power loss in megawatts.

Real power generation limit

$$P_{i\min} \leq P_i \leq P_{i\max} \quad (3)$$

$P_{i\min}$ is the minimum limit and $P_{i\max}$ is the maximum limit of real power of the i th unit in megawatts.

The inclusion of nonlinearities to above ELD problem is considered by two different types of NSELD problems.

2.2 ELD with Valve Point Loading (ELDVPL) [9]

The complex formula for the objective function of ELD with “valve point loading” is represented as below:

$$F_T = \min \left(\sum a_i P_i^2 + b_i P_i + c_i + |e_i \times \text{Sin}\{f_i \times (P_i^{\min} - P_i)\}| \right) (\$) \quad (4)$$

where e_i , f_i are constants from the valve point effect of the generating unit i . The objective function of ELDVPL is to minimize F_T of (4) subject to same set of constraints given in (2) and (3).

2.3 ELD with Prohibited Operating Zone and Ramp Rate Limit (ELDPOZRR) [4], [6]

The objective function of the ELDPOZRR problem is same as in (1) and in addition to the real power balance and generation limits which is same as in (2) and (3), two other constraints are also considered.

Generating unit ramp rate limits

$$\max(P_i^{\min}, P_{i0} - DR_i) \leq P_i \leq \min(P_i^{\max}, P_{i0} + UR_i) \tag{5}$$

where P_{i0} is the previous power output of the i th generating unit, UR_i / DR_i is the Up/Down ramp limits of generator i in megawatts.

Prohibited operating zone

The feasible operating zones of a unit can be described as follows:

$$P_i \in \begin{cases} P_{i,\min} \leq P_i \leq P_{i,1} \\ P_{i,k-1}^u \leq P_i \leq P_{i,k}^l, k = 2,3,\dots, pz_i \\ P_{i,pz_i}^u \leq P_i \leq P_{i,\max}, i = 1,2,\dots, n_{pz} \end{cases} \tag{6}$$

where $P_{i,k}^l / P_{i,k}^u$ are the lower/upper bounds of the prohibited operating zone of unit i , pz_i is the number of prohibited operating zones of unit i and n_{pz} is the number of units having prohibited operating zones.

3 NSO Tuned DE

The NSO tuned DE follows a new search technique by employing neighborhood of each individual. The neighborhood is selected for highly fitted individuals, which increases the exploitation ability of the DE for the ELD problem. The individual with low probability of being selected is abandoned from forming a highly fitted population. The abandoned individual is replaced by a newly generated individual. The generation of new individual is done randomly from the given search space of the problem. This process of abandoning and replacing a poor individual from forming a population enhances the exploration ability of DE. Thus each solution obtained by DE is subjected to fine tuning and the technique is directed towards the path of reaching global optimum. The objective of introducing the neighborhood search and the replacement criteria is to maintain a balance between exploration and exploitation ability of the DE. This would guarantee the DE to bypass the stagnation point and converges to a global peak point.

3.1 Algorithm for Proposed Method

Step 1. Get the data of the system

Step 2. Set the generation $G = 0$. Generate an initial population of NP D dimensional

vectors, $x_{i,G}$ using $x_{ij,G} = x_{ij}^{\min} + rand(0,1) \cdot (x_{ij}^{\max} - x_{ij}^{\min})$

where $i=1, 2, 3, \dots, NP$, $j=1, 2, \dots, D$, NP is the size of population, D is dimension of the problem and $rand(0,1)$ represents a uniformly distributed random variable within the range $[0, 1]$. x_{ij}^{\min} and x_{ij}^{\max} are the lower and upper bound of the search space for a given problem.

Step 3. Set iteration count, $iter = 1$. Evaluate each individual of the above population, $f(x_i)$

Step 4. Prepare a mutant vector for each target vector, $V_{i,G+1}$ using

$$V_i = x_{r1} + F \cdot (x_{r2} - x_{r3})$$

Step 5. Evaluate each individual of target vector, $f(v_i)$

Step 6. Perform a greedy selection between two vectors based on their fitness value.

Step 7. A new vector $u_{i,G+1}$ is created based on the fitness value of the two vectors

using, $u_{i,G+1} = \begin{cases} v_{i,G+1}, & \text{if } f(v_i) \geq f(x_i) \\ x_i, & \text{otherwise} \end{cases}$.

Step 8. Each individual member of $u_{i,G+1}$ is subjected to following equation,

$$P_i = \frac{fit_i}{\sum_{i=1}^{NP} fit_i}$$

where fit_i is the fitness of the i th individual above population and NP is the size of the population.

Step 9. Based on the P_i value of the i th individual a neighborhood vector is created

n_{ij} using $n_{ij} = x_{ij} + \phi(x_{ij} - x_{kj})$

where $j=1, 2, \dots, D$ and $k = \text{int}(\text{rand} * \text{NP}) + 1$ are randomly chosen indexes, ϕ is a random number between $[-1, 1]$.

Step 10. Individuals with poor value of P_i is replaced by new random vectors X_{new}

using $X_{new} = rand(0,1)(x_{hi} - x_{li})$.

Step 11. $iter = iter + 1$.

Step 12. until termination is reached

End

4 Simulation Results

To validate the effectiveness of the proposed method, it is tested with two test cases. The two test cases are 6 generator system and 13 generator system with non convex cost function. All of the test cases are simulated for 50 trial runs to validate the superiority and robustness of the proposed method compared to the other optimization methods. The result obtained from proposed method has been compared with DEC-SQP [11] for 13 generator system; with NPSO-LRS and other techniques for 6 generator system. The software has been written in MATLAB-R2009a language and executed on a 2.0-GHz Pentium Dual personal computer with 1400-MB RAM. In these case studies, the population size N was 50, F was 0.8 and the stopping criterion G_{max} was 500 generations for the NSO tuned DE. All the B coefficients are given in per unit (p.u.) on a 100 MVA base capacity.

Case study 1

This test case is a NSELD without considering the transmission losses, ramp rate limits and prohibited operating zones. This test system consists of 13 units with valve point loading and has a load demand of 1800MW. The input data are given in [11]. The result obtained from proposed method NSO tuned DE has been compared with DEC-SQP [11] and other methods. Table I shows the frequency of convergence in 50 trial runs. It is clear from Table 1 and 2 that the proposed method produces a much better solution with less computation time compared to IPM-DE, STHDE and other methods. It can also be seen that the average fuel cost produced by the NSO tuned DE is less compared to other methods.

Table 1. Comparison among Different Methods after 50 Trials (13 unit system)

Method	Total Generation Cost (\$)		
	Minimum Cost	Mean Cost	Maximum Cost
NSO-DE	16413.357	16537.962	16702.866
DEC(1)-SQP(1)	17938.9521	17943.1339	17944.8105
IPM-DE	17940.89	17950.42	17961.2314
STHDE	NA	NA	NA
ICA-PSO	17960.37	17967.94	17978.14
IGA_MU	17963.9848	NA	NA
PS	17969.17	18233.52	18088.84

Table 2. Frequency of Convergence for 13 unit System

Method	Range of Cost (\$)			
	16400-16500	16500-16600	16600-16700	16700-16800
NSO-DE	12	24	11	3

Case study 2

A system with six generators with ramp rate limit and prohibited operating zone is used here and has a total load of 1263 MW. The input data have been adopted from [12]. Results obtained from DE, proposed NSO tuned DE, PSO and new coding-based modified PSO [13] and other methods have been presented here. Table 3 shows the frequency of convergence in 50 trial runs. It is clear from Table 3 and 4 that the proposed method produces a much better solution with less computation time compared to other methods. It can also be seen that the average fuel cost produced by the NSO tuned DE is less compared to PSO, IDE and other methods.

Table 3. Comparison among Different Methods after 50 Trials (6 unit System)

Method	Total Generation Cost (\$)		
	Minimum Cost	Mean Cost	Maximum cost
NSO tuned DE	15,286	15,294	15,310
IDE	15,351	15,356	15,359
PSO	15,450	15,454	15,492
GA	15,459	15,469	15,469

Table 4. Frequency of Convergence for 6 unit system

Method	Range of Cost (\$)			
	15000-15300	15300-15500	15500-15800	15800-16000
NSO-DE	26	12	9	3

5 Conclusion

A hybrid approach by fine tuning Differential evolution algorithm using Neighborhood Search Operator for solving the Non-smooth economic load dispatch problem of units with valve-point effects is presented. An optimal range of mutation rate, probability factor, and replacement size for the NSO tuned DE algorithm is estimated to solve all the test cases considered in this paper. The feasibility of the NSO tuned DE method was illustrated by conducting case studies on a 6 unit and 13 unit systems with valve-point effects and ramp rate limits and compared with the results obtained using the other methods. In each test case, the quality of solution, convergence property and reliability demonstrates the superiority of the proposed NSO tuned DE method over other optimization techniques for solving the NSELD problem.

References

1. Abido, M.A.: A novel multiobjective evolutionary algorithm for environmental/economical power dispatch. *Elect. Power Syst. Res.* 65, 71–81 (2003)
2. Hindi, K.S., Ab Ghani, M.R.: Dynamic economic dispatch for large scale power systems: a Lagrangian relaxation approach. *Elect. Power Syst. Res.* 13(1), 51–56 (1991)

3. Wood, A.J., Wollenberg, B.F.: *Power Generation, Operation and Control*, 2nd edn. Wiley, New York (1996)
4. Li, F., Morgan, R., Williams, D.: Hybrid genetic approaches to ramping rate constrained dynamic economic dispatch. *Elect. Power Syst. Res.* 43(2), 97–103 (1997)
5. Attaviriyanupap, D., Kita, H., Tanaka, E., Hasegawa, J.: A hybrid EP and SQP for dynamic economic dispatch with nonsmooth incremental fuel cost function. *IEEE Trans. Power Syst.* 17(2), 411–416 (2002)
6. Lee, F.N., Arthur, M.B.: Reserve constrained economic dispatch with prohibited operating zones. *IEEE Trans. Power Syst.* 8(1), 246–253 (1993)
7. Yang, H.T., Yang, P.C., Huang, C.L.: Evolutionary programming based economic dispatch for units with nonsmooth incremental fuel cost functions. *IEEE Trans. Power Syst.* 11(1), 112–118 (1996)
8. Wong, K.P., Wong, Y.W.: Genetic and genetic/simulated-annealing approaches to economic dispatch. *Proc. Inst. Elect. Eng., Gener. Transm. Distrib.* 141(5), 507–513 (1994)
9. Victoire, T.A.A., Jeyakumar, A.E.: Hybrid PSO-SQP for economic dispatch with valve-point effect. *Elect. Power Syst. Res.* 71(1), 51–59 (2004)
10. Storn, R., Price, K.: *Differential Evolution—A Simple and Efficient Adaptive Scheme for Global Optimization Over Continuous Spaces*. International Computer Science Institute, Berkeley, CA, Tech. Rep. TR-95-012 (1995)
11. dos Santos Coelho, L., Mariani, V.C.: Combining of chaotic differential evolution and quadratic programming for economic dispatch optimization with valve-point effect. *IEEE Trans. Power Syst.* 21(2), 989–996 (2006)
12. He, D., Wang, F., Mao, Z.: A Hybrid genetic algorithm approach based on differential evolution for economic dispatch with valve point effect. *Int. J. Elect. Power Energy Syst.* 30(1), 31–38 (2008)
13. Kuo, C.C.: A novel coding scheme for practical economic dispatch by modified particle swarm approach. *IEEE Trans. Power Syst.* 23(4), 1825 (2008)

Performance Evaluation of Noise Subspace Methods of Frequency Estimation Techniques

Kakaraparathi Bramaramba¹, S. Koteswara Rao², and K. Raja Rajeswari³

¹ ECE Department, A.U

² NSTL

³ ECE Department, A.U, Visakhapatnam, India
bramarambak@gmail.com

Abstract. Frequency Estimation methods have the ability to resolve complex exponentials that are closely spaced in frequency. The estimation of the frequencies is based on the eigen decomposition of the autocorrelation matrix of the input data. The autocorrelation matrix after eigen decomposition produces two subspaces, namely noise subspace and signal subspace. The methods that are based on the estimation of frequencies using noise subspace of the autocorrelation matrix are called Noise subspace methods of Frequency Estimation. Pisarenko Harmonic Decomposition, MUSIC method, Eigen Vector method and the Minimum Norm methods belongs to the category of Noise subspace methods. This paper investigates the performance evaluation of all the Noise Subspace methods of frequency estimation techniques for a common Synthetic Power signal having harmonics at 600Hz, 900Hz and 1500Hz with a sampling frequency of 3000Hz. Extensive Monte-Carlo simulation is carried out for ten numbers of times and the simulated figures are shown. The values obtained after the application of Noise subspace methods are compared with that of the actual inputs and are tabulated. The simulation of all methods is performed by using MATLAB software.

Keywords: Autocorrelation matrix, Eigen decomposition, Eigen Vector method, Minimum Norm method, MUSIC method, Noise Subspace, Pisarenko Harmonic Decomposition.

1 Introduction

The methods of Spectrum Estimation which have the ability to resolve complex exponentials that are closely spaced in frequency are known as Harmonic or Frequency Estimation methods [1, 2]. These methods use models in estimating the power spectrum of a WSS random process. The estimation of frequencies depends on the eigen decomposition of the autocorrelation matrix into subspaces, a signal subspace and a noise subspace. The Pisarenko Harmonic Decomposition method, MULTI SIGNAL Classification (MUSIC) method, Eigen Vector method and Minimum Norm method belongs to the category of Noise subspace methods of frequency estimation. Section 1.1 describes the eigen decomposition of the autocorrelation matrix.

In order to motivate the use of an eigendecomposition of the autocorrelation matrix as an approach that may be used for frequency estimation, consider the first-order process

$$x(n) = A_1 e^{jn\omega_1} + \omega(n) \tag{1}$$

That consists of a single complex exponential in white noise. The amplitude of the complex exponential is $A_1 = |A_1| e^{j\phi_1}$ where ϕ_1 is a uniformly distributed random variable, and $\omega(n)$ is white noise that has a variance of σ_ω^2 , the autocorrelation sequence of $x(n)$ is

$$r_x(k) = P_1 e^{jk\omega_1} + \sigma_\omega^2 \delta(k) \tag{2}$$

where $P_1 = |A_1|^2$ is the power in the complex exponential. Therefore, the $M \times M$ autocorrelation matrix for $x(n)$ is a sum of an autocorrelation matrix due to the signal, \mathbf{R}_s , and an autocorrelation matrix due to the noise, \mathbf{R}_n ,

$$\mathbf{R}_x = \mathbf{R}_s + \mathbf{R}_n \tag{3}$$

It is possible to extract all of the parameters of interest about $x(n)$ from the eigenvalues and eigenvectors of \mathbf{R}_x as follows:

1. Perform an eigendecomposition of the autocorrelation matrix, \mathbf{R}_x . The largest eigenvalue will be equal to $MP_1 + \sigma_w^2$ and the remaining eigenvalues will be equal to σ_w^2 .
2. Use the eigenvalues of \mathbf{R}_x to solve for the power P_1 and the noise variance as follows:

$$\begin{aligned} \sigma_w^2 &= \lambda_{\min} \\ P_1 &= \frac{1}{M} (\lambda_{\max} - \lambda_{\min}) \end{aligned} \tag{4}$$

3. Determine the frequency ω_1 from the eigenvector \mathbf{v}_{\max} that is associated with the largest eigenvalue using, for example, the second coefficient of \mathbf{v}_{\max} ,

$$\omega_i = \arg\{v_{\max}(1)\} \tag{5}$$

2 Mathematical Modeling

The Frequency Estimation methods use models in estimating the power spectrum of a WSS random process. Various models are used for estimating the frequencies of complex exponentials in noise using the noise subspace of the eigen decomposed

autocorrelation matrix. The following sections give the detailed mathematical modeling of the noise subspace methods of frequency estimation techniques.

2.1 Pisarenko Harmonic Decomposition

This method is based on the determination of frequencies that are derived from the eigenvector corresponding to the minimum eigenvalue of the autocorrelation matrix. The steps involved in the determination of frequencies using Pisarenko Harmonic Decomposition method are summarized as follows:

Step 1: Given that a process consists of p complex exponentials in white noise, find the minimum eigenvalue λ_{\min} and the corresponding eigenvector \mathbf{v}_{\min} of the $(p + 1) \times (p + 1)$ autocorrelation matrix \mathbf{R}_x .

Step 2: Set the white noise power equal to the minimum eigenvalue, $\lambda_{\min} = \sigma_w^2$, and set the frequencies equal to the angles of the roots of the eigenfilter

$$V_{\min}(z) = \sum_{k=0}^p v_{\min}(k)z^{-k} \tag{6}$$

or the location of the peaks in the frequency estimation function

$$\hat{P}_{PHD}(e^{j\omega}) = \frac{1}{|\mathbf{e}^H \mathbf{v}_{\min}|^2} \tag{7}$$

Step 3: Compute the powers of the complex exponentials by solving the linear equations (8).

$$\begin{bmatrix} |V_1(e^{j\omega_1})|^2 & |V_1(e^{j\omega_2})|^2 & |V_1(e^{j\omega_3})|^2 & |V_1(e^{j\omega_4})|^2 \\ |V_2(e^{j\omega_1})|^2 & |V_2(e^{j\omega_2})|^2 & |V_2(e^{j\omega_3})|^2 & |V_2(e^{j\omega_4})|^2 \\ \vdots & \vdots & \vdots & \vdots \\ |V_p(e^{j\omega_1})|^2 & |V_p(e^{j\omega_2})|^2 & \dots & |V_p(e^{j\omega_4})|^2 \end{bmatrix} \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_p \end{bmatrix} = \begin{bmatrix} \lambda_1 - \sigma_w^2 \\ \lambda_2 - \sigma_w^2 \\ \vdots \\ \lambda_p - \sigma_w^2 \end{bmatrix} \tag{8}$$

2.2 MUSIC Method

This method determines the frequencies of complex exponentials in noise by reducing the effects of spurious peaks. To see how the MUSIC algorithm works, assume that $x(n)$ is a random process consisting of p complex exponentials in white noise with a variance of σ_w^2 , and let \mathbf{R}_x be the $M \times M$ autocorrelation matrix with $M > p + 1$. If the eigenvalues of \mathbf{R}_x are arranged in decreasing order, $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_M$, and if $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_M$ are the corresponding eigenvectors, then we may divide these eigenvectors into two groups: the p signal eigenvectors corresponding to the p largest eigenvalues, and the $M - p$ noise eigenvectors that, ideally, have eigenvalues equal to σ_w^2 .

Although we could consider estimating the white noise variance by averaging the $M-p$ smallest eigenvalues

$$\sigma_w^2 = \frac{1}{M - p} \sum_{k=p+1}^M \lambda_k \tag{9}$$

Estimating the frequencies of the complex exponentials is a bit more difficult. Since the eigenvectors of \mathbf{R}_x are of length M , each of the noise subspace eigenfilters

$$V_i(z) = \sum_{k=0}^{M-1} v_i(k) z^{-k}; \quad i = p+1, \dots, M \tag{10}$$

will have $M-1$ roots (zeros). Ideally, p of these roots will lie on the unit circle at the frequencies of the complex exponentials, and the eigen spectrum

$$\left| V_i(e^{j\omega}) \right|^2 = \frac{1}{\left| \sum_{k=0}^{M-1} v_i(k) e^{-jk\omega} \right|^2} \tag{11}$$

associated with the noise eigenvector \mathbf{V}_i will exhibit sharp peaks at the frequencies of the complex exponentials. However, the remaining $(M-p-1)$ zeros may lie anywhere and, infact, some may lie close to the unit circle, giving rise to spurious peaks In the eigenspectrum. Furthermore, with inexact autocorrelations, the zeros of $V_i(z)$ that are on the unit circle may not remain on the unit circle. Therefore, when only one noise eigenvector is used to estimate the complex exponential frequencies, there may be some ambiguity in distinguishing the desired peaks from the spurious ones[1,4]. In the MUSIC algorithm, the effects of these spurious peaks are reduced by averaging, using the frequency estimation function

$$\hat{P}_{MU}(e^{j\omega}) = \frac{1}{\sum_{i=p+1}^M \left| \mathbf{e}^H \mathbf{v}_i \right|^2} \tag{12}$$

The frequencies of the complex exponentials are then taken as the locations of the p largest peaks in $\hat{P}_{MU}(e^{j\omega})$. Once the frequencies have been determined the power of each complex exponential may be found using Eq.(12).

2.3 Eigen Vector Method

The Frequency Estimation method in which the frequencies of complex exponentials in noise are determined by reducing the effects of spurious peaks by averaging and this procedure also involves multiplication of the inverse of eigenvalues associated with the eigen vectors is known as Eigen Vector method. Specifically, the EV method estimates the exponential frequencies from the peaks of the eigenspectrum

$$\hat{P}_{EV}(e^{j\omega}) = \frac{1}{\sum_{i=p+1}^M \frac{1}{\lambda_i} |\mathbf{e}^H \mathbf{v}_i|^2} \tag{13}$$

where λ_i is the eigenvalue associated with the eigenvector \mathbf{v}_i

2.4 Minimum Norm Method

The minimum norm algorithm uses a single vector \mathbf{a} that is constrained to lie in the noise subspace, and the complex exponential frequencies are estimated from the peaks of the frequency estimation function,

$$\hat{P}_{MN}(e^{j\omega}) = \frac{1}{|\mathbf{e}^H \mathbf{a}|^2} \tag{14}$$

With \mathbf{a} constrained to lie in the noise subspace, if the autocorrelation sequence is known exactly, then $|\mathbf{e}^H \mathbf{a}|^2$ will have nulls at the frequencies of each complex exponential. Therefore, the z-transform of the coefficients in \mathbf{a} may be factored as follows:

$$A(z) = \sum_{k=0}^{M-1} a(k)z^{-k} = \prod_{k=1}^p (1 - e^{j\omega_k} z^{-1}) \prod_{k=p+1}^{M-1} (1 - z_k z^{-1}) \tag{15}$$

where z_k for $k = p+1, \dots, M-1$ are the spurious roots that do not, in general, lie on the unit circle. The problem then is to determine which vector in the noise subspace minimizes the effects of the spurious zeros on the peaks of $\hat{P}_{MN}(e^{j\omega})$. The approach that is used in the minimum norm algorithm is to find the vector \mathbf{a} that satisfies the following three constraints:

1. The vector \mathbf{a} lies in the noise subspace.
2. The vector \mathbf{a} has minimum norm.
3. The first element of \mathbf{a} is unity.

3 Selection Criteria for Performance Evaluation

An important factor in the selection of a spectrum estimation technique is the performance of the estimator. In comparing one non-parametric method to another, there is a trade-off between resolution and variance. The variability, ν of the estimate is represented as,

$$\nu = \frac{\text{var} \left\{ \hat{P}_x(e^{j\omega}) \right\}}{E^2 \left\{ \hat{P}_x(e^{j\omega}) \right\}} \tag{16}$$

The variability must be as low as possible in order to determine the given non-parametric method as the best method.

Resolution, Δw of the estimate is represented as,

$$\Delta w = f_2 - f_1 \tag{17}$$

where $f_2 - f_1$ is the bandwidth of the mainlobe [4,5].

The resolution must be high in order to determine the given non-parametric method as the best method.

The overall figure of merit μ is defined as the product of the variability, ν and the resolution Δw .

$$\mu = \nu \Delta w \tag{18}$$

As the figure of merit decreases the performance of the non-parametric method increases, so the figure of merit should be as low as possible [6].

4 Monte-Carlo Simulation of a Synthetic Signal Consisting of Harmonics

For the purpose of simulation a signal $x(n)$ consisting of three complex exponentials in white noise is considered. It is represented as,

$$x(n) = \sum_{k=1}^3 A_k e^{j(n\omega_k + \phi_k)} + w(n) \tag{19}$$

where the amplitudes A_k are equal to one, the frequencies ω_k are 0.2π , 0.3π and 0.5π (the denormalized frequencies are 200Hz, 300Hz and 500Hz), the phases are uncorrelated random variables that are uniformly distributed over the interval $[0, 2\pi]$, and the variance of the white noise is $\sigma_w^2 = 0.5$. Using ten different realizations of $x(n)$ with $N = 64$ values, overlay plots of the frequency estimation functions using Pisarenko’s method, the MUSIC algorithm, the eigenvector method, and the minimum norm algorithm are shown in the Figs 1(a), 2(a), 3(a) and 4(a) respectively. The average of the Monte-Carlo simulated plots are shown in Fig.s 1(b), 2(b), 3(b) and 4(b) respectively.

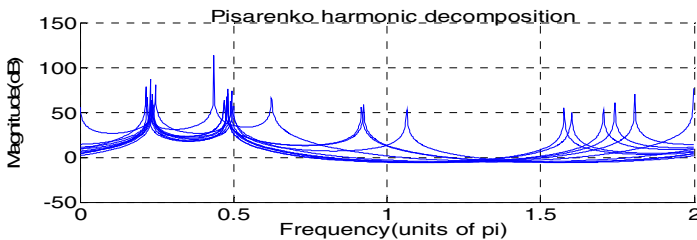


Fig. 1(a). Monte-Carlo simulated Pisarenko’s estimates of $x(n)$

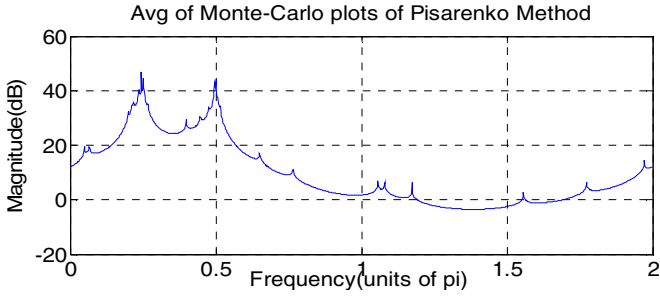


Fig. 1(b). Avg of Monte-Carlo simulated Pisarenko's estimates of $x(n)$

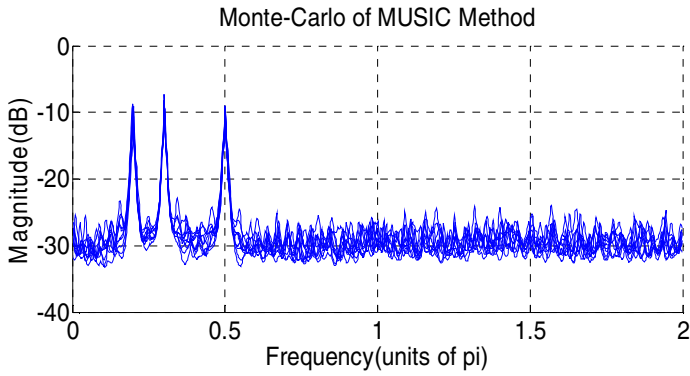


Fig. 2(a). Monte-Carlo simulated MUSIC estimates of $x(n)$

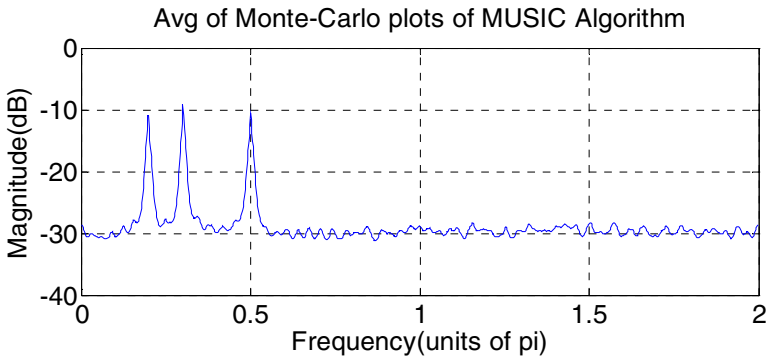


Fig. 2(b). Avg of Monte-Carlo simulated MUSIC estimates of $x(n)$

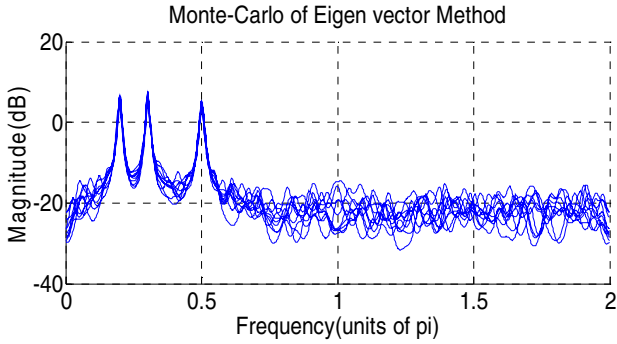


Fig. 3(a). Monte-Carlo simulated Eigen Vector estimates of $x(n)$

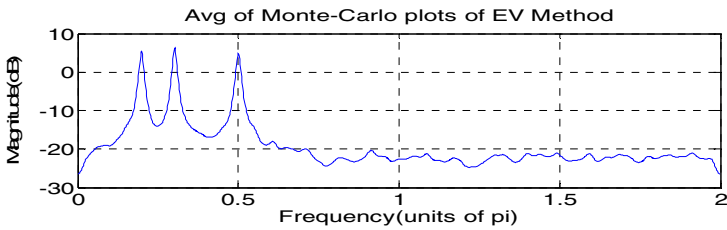


Fig. 3(b). Avg of Monte-Carlo simulated Eigen Vector estimates of $x(n)$

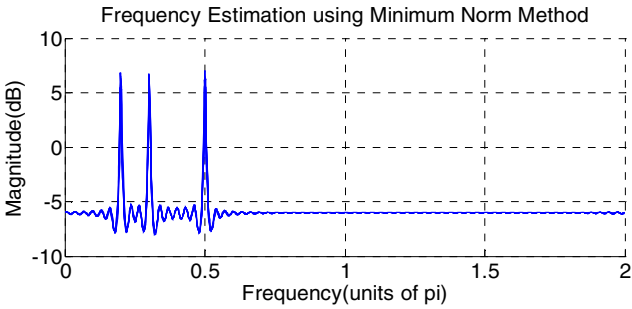


Fig. 4(a). Monte-Carlo simulated Minimum Norm estimates of $x(n)$

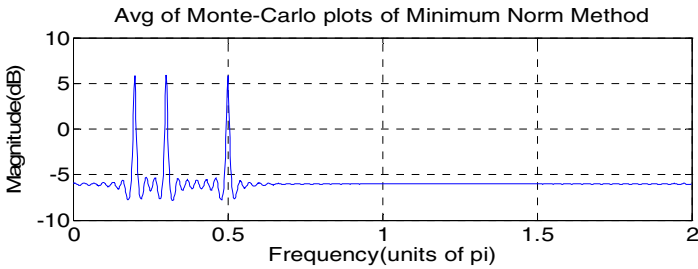


Fig. 4(b). Avg of Monte-Carlo simulated Minimum Norm estimates of $x(n)$

The estimated frequencies are compared with the true values and are represented in Table 1.

Table 1. Comparison of Estimated Vs True Frequencies for Various Noise Subspace based Methods

Method Used	Frequency (Hz)		Frequency (Hz)		Frequency (Hz)	
	True	Estimated	True	Estimated	True	Estimated
Pisarenko Harmonic Decomposition	600	732.3000	900	750	1500	1500
MUSIC	600	597.6000	900	908.4000	1500	1500
Eigen Vector	600	591.9000	900	906.1000	1500	1500
Minimum Norm	600	597.6000	900	902.4000	1500	1500

Performance Evaluation of Noise subspace methods of Frequency Estimation is done according to the selection criteria and is represented in Table 2. Since the estimation of frequencies can be compared with the true values, one more parameter named accuracy is added in the evaluation process which gives the closeness of the estimated result to the true value.

Table 2. Performance Evaluation of Noise subspace based Frequency Estimation methods

Method Used	Variability	Resolution	Figure of merit	Accuracy
Pisarenko	1.4876	0.6	0.8926	33.33%
MUSIC	0.0108	0.9	0.0097	92%
Eigen Vector	0.0901	0.75	0.0676	92%
Minimum Norm	0.0471	0.7	0.0330	94%

5 Conclusion

A synthetic power signal having harmonics at 600Hz, 900Hz and 1500Hz with a sampling frequency of 3000Hz is simulated using extensive Monte-Carlo simulation

for ten times. It is observed from the simulated results and Tabular forms 1.0 and 2.0 that the performance of MUSIC method is best when compared to all other Noise subspace based Frequency Estimation techniques, as it produced least variability, figure of merit, good accuracy and highest resolution. It is also observed from the simulated results that the effect of spurious peaks which gives ambiguity regarding the detection of exact harmonic frequencies is least with the MUSIC method and highest with the Pisarenko Harmonic Decomposition. Therefore, the MUSIC method exactly suits in predicting the presence of harmonic frequencies as well as magnitudes.

References

- [1] Hayes, M.H.: Statistical Digital Signal Processing and Modeling. John Wiley & Sons, INC. (1996)
- [2] Proakis, J.G., Manolakis, D.G.: Digital Signal Processing Principles, Algorithms and Applications. PHI (2002)
- [3] Bollen, M.H.J., Gu, I.Y.H.: Signal Processing of Power Quality Disturbances. IEEE Press Series on Power Engineering (2011)
- [4] Kay, S., Marple Jr., S.L.: Sources and remedies for spectral line splitting in autoregressive spectrum analysis. In: Proc. Int. Conf. on Acoust, Speech, Sig. (1979)
- [5] Chapman, S.J.: MATLAB programming for Engineers. Thomson Pub., Toronto (2008)
- [6] Elliott, D.F.: Handbook of Digital Signal Processing Engineering Applications. Rockwell International Corporation Pub., Anaheim (2006)

Performance Analysis of DSSS System Using Adaptive Filters in Interference Prone Environment

Katyayani Kaligathi¹, Kandula Srinivasa Rao¹,
Seetala Santha Kumari², and G. Prabhakara Rao³

¹G.V.P College of Engineering for Women, E.C.E, Vizag.

²Andhra University, E.C.E, vizag.

³J.N.T University, E.C.E, Kakinada

{klgthkati,ksrinivas.ece}@gmail.com, santakseetala@yahoo.com,
director_evaluation@jntuk.edu.in

Abstract. Direct Sequence Spread Spectrum (DSSS) communication techniques offer a promising solution to an overcrowded spectrum amid growing demand for mobile and personal communication services. In mobile and personal communication systems, DSSS systems share the same frequency band with the existing narrow band frequency communication system. Therefore the performance would inevitably be affected by narrow band interference (NBI). In this paper, we present an interference suppression method in Direct Sequence Spread Spectrum (DSSS) systems using adaptive filters. Demonstrated by MATLAB simulation results, the method significantly improves the performance of DSSS receiver serving at the narrow band interference environment. The improvement of resulting BER after NBI suppression will further enhance along with the increasing of interference intensities. Compared with Least Mean Squares adaptive (LMS) algorithm, Recursive Least Squares (RLS) adaptive algorithm has better real time performance.

Keywords: Spread spectrum, Adaptive filters, Narrow Band Interference (NBI), Least Mean Squares (LMS), Recursive Least Squares (RLS) algorithm, Bit Error Rate (BER).

1 Introduction

As the wireless personal communications field has grown over the last few years, the method of communication known as spread spectrum has gained a great deal of prominence. Spread spectrum involves spreading the desired signal over a bandwidth much larger than the minimum bandwidth necessary to send the signal. Spread spectrum systems afford protection against jamming (intentional interference) and interference from other users in the same band as well as noise by “spreading” the signal to be transmitted and performing the reverse “de-spread” operation on the received signal at the receiver. This de-spreading operation in turn spreads those signals which are not properly spread when transmitted, decreasing the effect that spurious signals will have on the desired signal. Spread Spectrum systems can be

thought of as having two general properties: first, they spread the desired signal over a bandwidth much larger than the minimum bandwidth needed to send the signal, and secondly this spreading is carried out using a Gold sequence which is generated using two pseudorandom noise (PN) sequences. There are two fundamental techniques for spectrum spreading: Direct sequence spread spectrum (DSSS) and Frequency hopping spread spectrum (FHSS). Direct sequence spread spectrum combines the information signal with a spreading signal having much wider bandwidth. The net modulation signal effectively handles the wide bandwidth of the spreading signal. This wide modulation is then applied to a fixed frequency carrier signal for transmission. The spreading code directly spreads the information, ahead and independent of the RF modulator. Frequency hopping takes the opposite approach. Rather than spreading the modulation about a fixed carrier, the information is left unchanged and the spreading signal is used to change the frequency of the carrier provided by the carrier generator. The data directly modulate the hopping carrier. One of the major features that distinguish modern wireless communication channels is the significant amount of structured interference that must be contended with in wireless channels. This interference is inherent in many wireless systems due to their operation as *multiple-access* systems, in which multiple transmitter / receiver pairs communicate through the same physical channel using non-orthogonal multiplexing. Structured interference also arises because of other non-systemic features of wireless systems, such as the desire to share bandwidth with other, dissimilar, communication services. Signal processing plays a central role in suppression of the structured interference arising in wireless communication systems. In particular, the use of appropriate signal processing methods can make a significant difference in the performance of such systems. Moreover, since many wireless systems operate under highly dynamic conditions because of the mobility of the trans-receivers and of the random nature of the channel access, *adaptive* signal processing is paramount in this context. The study of adaptive processing techniques for interference suppression in wireless systems has been a very active area of research in recent years. This presentation focuses primarily on the problem of Narrow Band Interference (NBI) suppression, which limits source of interference for the spread spectrum communications. In this work, LMS and RLS adaptive algorithms are employed to achieve the reduced BER after Narrow Band Interference filtering. Simulation results show that RLS provides better interference suppression and significant reduction in BER. There is abrupt change in interference suppression and bit error rate improvement when compared to [1].

This paper is organized as follows: In section 2&3, DSSS Transmitter model and Interference suppression receiver model are presented, in section 4, Adaptive algorithms LMS and RLS are explained in detail. Simulation results such as spectrum of DSSS signal before and after interference suppression and BER plot are provided in section 5 and finally conclusions are given in section 6.

2 DSSS Transmitter Model

The DSSS system transmitter module is shown in fig1.

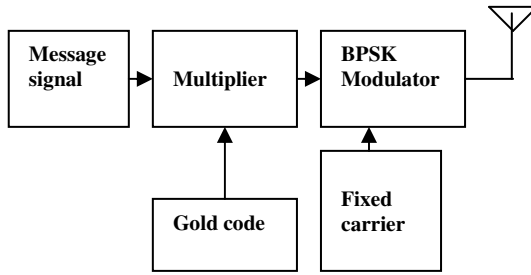


Fig. 1. Block diagram of DSSS transmitter

The information bearing baseband signal is denoted as $v(t)$ and is expressed as

$$v(t) = \sum_{n=-\infty}^{+\infty} a_n g_T(t - nT_b) \tag{1}$$

Where $a_n=0$ or 1 , $-\infty < n < +\infty$ and $g_T(t)$ is a rectangular pulse train of duration T_b . This signal is multiplied with the signal from Gold code sequence generator which may be expressed as

$$c(t) = \sum_{n=-\infty}^{+\infty} c_n p(t - nT_c) \tag{2}$$

Where c_n represents the binary Gold code sequence of 0's or 1's, $p(t)$ is a rectangular pulse of duration T_c . This multiplication operation serves to spread the bandwidth of the information bearing signal. The product signal $v(t)c(t)$ is used to modulate carrier $A_c \cos(2\pi f_c t + \theta)$ and generate BPSK signal,

$$u(t) = A_c \cos(2\pi f_c t + \theta) \tag{3}$$

where A_c is 1. The carrier signal is transmitted with same phase ($\theta=0^\circ$) when $v(t)c(t)=0$ and is shifted by 180° ($\theta=180^\circ$) when $v(t)c(t)=1$. The rectangular pulse $p(t)$ is usually called a chip and its time duration T_c is called the chip interval. The reciprocal $1/T_c$ is called the chip rate and corresponds to the bandwidth of the transmitted signal.

3 Interference Suppression

The performance of DSSS systems are affected by narrow band interference (NBI). Since NBI becomes a wideband signal after de spreading, the system inherently can suppress NBI to a certain extent by using normal filter. However, the comparable performance is based on the assumption that direct spread (DS) signal power is more or the spectrum of DS signal is much wider than the NBI. In case of the spreading gain decreases or interference signal is strong, this in turn may cause the performance to be somewhat unsatisfactory. Due to the complex radio environment and the continuously changing of the interference signal, the statistical characteristics are always changing which leads to the fact that the algorithms of fixed coefficients cannot effectively suppress the time-varying narrow band interference. Hence an adaptive filter which automatically adjusts the filter coefficients according to changing interference signal is used to suppress the interference. Figure2 shows the suppression model.

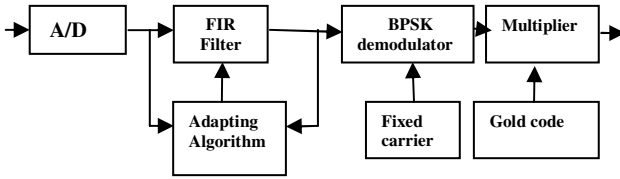


Fig. 2. Suppression receiver model

The received signal can be expressed as

$$S(t)=d(t)+n(t)+i(t) \tag{4}$$

where $d(t)$ is modulated DS-SS signal $n(t)$ is additive white Gaussian noise with power spectral density $N_0/2$, $i(t)$ is narrow band interference which can be represented as finite sum of sinusoids and expressed as

$$i(t) = \sum_{j=1}^K A_j \cos(2\pi f_j t + \phi_j) \tag{5}$$

where K is the number of single frequency interferences. A_j is the amplitude of single frequency interference. f_j is the frequency of single frequency interference. ϕ_j is the original phase of single frequency interference. The received modulated DS-SS which contains narrow band interference and white Gaussian noise is sampled by A/D first and then it is filtered using adaptive algorithms and then demodulated to get the DS-SS signal. Then the demodulated signal is multiplied with a replica of waveform $c(t)$ generated by gold code sequence generator at the receiver which is synchronized to the gold code in the received signal. This operation is called despreading since the effect of multiplication by $c(t)$ at the receiver is to undo the spreading. The threshold detector detects the transmitted bit by comparing each sample value with the threshold generated by it.

4 Adaptive Algorithms

An adaptive FIR or IIR filter design itself based on the characteristics of the input signal to the filter and a signal which represent the desired behavior of the filter on its input. Designing the filter does not require any other frequency response information or specification. The adaptive algorithm is used to reduce the error between the output signal $y(n)$ and the desired signal $d(n)$. There are two types of adapting algorithms

- 1) Least Mean Squares (LMS) and
- 2) Recursive Least Squares (RLS)

Figure 3 shows the block diagram of adaptive filter

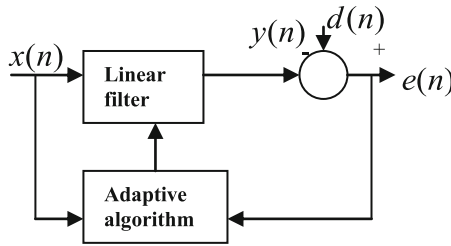


Fig. 3. Adaptive filter

A. LMS Filter

In LMS algorithm, the tap inputs $x(n), x(n-1), \dots, x(n-M+1)$ form the elements of the M by 1 tap input vector. Correspondingly the tap weights $w_0(n), w_1(n), \dots, w_{n-1}(n)$ form the elements of M by 1 tap weight vector $w(n)$. The value computed for this vector using LMS algorithm represents an estimate whose expected value may come close to the wiener solution as the number of iterations n approaches infinity. During the filtering process, the desired response $d(n)$ is supplied for processing, alongside the tap input vector $x(n)$. Given this input, the transversal filter produces an output which is used as an estimate of the desired response $d(n)$. The estimation error $e(n)$ and the tap input vector $x(n)$ are applied to the control mechanism and the feedback loop around the tap weights $w(n)$ is thereby closed. The scaling factor used in the computation of updating tap weights is denoted by a positive quantity μ called the step size parameter. By assigning a small value to μ , the adaptive process is made to progress slowly and the effects of gradient noise on the tap weights are largely filtered out. The updated value of the tap weight vector $w(n+1)$ according to the steepest descent algorithm is given by

$$w(n+1) = w(n) + \mu x(n) e^*(n) \tag{6}$$

where

$$e(n) = d(n) - y(n) = d(n) - w^H(n)x(n) \tag{7}$$

B. RLS Filter

In recursive implementations of the method of least squares, the computation is started with prescribed initial conditions and the information contained in the new data samples is used to update the old samples. The length of observable data is variable. The cost function is expressed to be minimized as $\varepsilon(n)$, where n is the variable length of observable data. A weighting factor is also introduced into the definition of $\varepsilon(n)$. The cost function is given by

$$\varepsilon(n) = \sum_{i=1}^n \beta(n,i) |e(i)|^2 \tag{8}$$

where $e(i)$ is the difference between the desired response $d(i)$ and the output $y(i)$ produced by a transversal filter whose tap inputs (at time i) equal to $x(i), x(i-1), \dots, x(i-M+1)$ where M is the length of the filter. That is,

$$e(i) = d(i) - y(i) = d(i) - w^H(n)x(i) \tag{9}$$

where $x(i)$ is the tap input vector at time i , defined by and $w(n)$ is the tap-weight vector at time n defined by

$$w(n)=[w_o(n),w_l(n),\dots,w_{M-1}(n)]^T \tag{10}$$

The tap weights of the transversal filter remain fixed during the observation interval $1 \leq i \leq n$ for which the cost function is defined. The weighting factor $\beta(n,i)$ has the property that

$$0 < \beta(n,i) \leq 1$$

A special form of weighting that is commonly used is the exponential weighting factor or forgetting factor defined by,

$$\beta(n,i) = \lambda^{n-i} \tag{11}$$

where λ is a positive constant close to but less than unity.

The cost function is expanded to be minimized as the sum of two components.

$$\varepsilon(n) = \sum_{k=1}^n \beta(n,i) |e(i)|^2 + \delta \lambda^n \|w(n)\|^2 \tag{12}$$

The two components of cost function are as follows:

1. The sum of weighted error squares,

$$\sum_{i=1}^n \lambda^{n-i} |e(i)|^2 = \sum_{i=1}^n \lambda^{n-i} |d(i) - w^H(n)x(i)|^2 \tag{13}$$

which is data dependent. This component measures the exponentially weighted error between the desired response $d(i)$ and the actual response of the filter $y(i)$ which is related to tap input vector $x(i)$ by the formula

$$y(i) = w^H(n)x(i) \tag{14}$$

where $w^H(n)$ is the Hermitian of tap weight vector $w(n)$

2. A regularizing term,

$$\delta \lambda^n \|w(n)\|^2 = \delta \lambda^n w^H(n)w(n) \tag{15}$$

where δ is a positive real number called the regularization parameter. Except for the factor $\delta \lambda^n$, the regularizing term depends solely on the tap weight vector $w(n)$. The term is included in the cost function to stabilize the solution to the recursive least squares problem by smoothing the solution.

The cost function $\varepsilon(n)$, is equivalent to reformulation of the M by M time average correlation matrix of the tap input vector $x(i)$ given by

$$\Phi(n) = \sum_{i=1}^n \lambda^{n-i} x(i)x^H(i) + \delta \lambda^n I \tag{16}$$

In this equation, I is the M by M identity matrix. The inverse of the correlation matrix is given by

$$P(n) = \Phi^{-1}(n) = \lambda^{-1} \Phi^{-1}(n-1) - \frac{\lambda^2 \Phi^{-1}(n-1)x(n)x^H(n)\Phi^{-1}(n-1)}{1 + \lambda^1 x^H(n)\Phi^{-1}(n-1)x(n)} \tag{17}$$

and gain vector $k(n)$ is given by

$$k(n) = \frac{\lambda^{-1} P(n-1)x(n)}{1 + \lambda^{-1} x^H(n)P(n-1)x(n)} \tag{18}$$

The apriori estimation error is

$$\zeta(n) = d(n) - x(n)w^H(n-1) \tag{19}$$

The time update for the tap weight vector is given by

$$w(n) = w(n-1) + k(n)\zeta^*(n) \tag{20}$$

5 Simulation Results

Simulation parameters are set as follows: Data rate is set to 127bps, 6 bit Gold code sequence of length 63 is used and spread spectrum code rate obtained is 8.001Kbps. The feedback polynomials used are $1+x^2+x^3$ and $1+x^3+x^4$. Single tone interference signal is of unity amplitude and frequency 16 KHz.

Fig.4 depicts the spectrum of interference affected DSSS signal

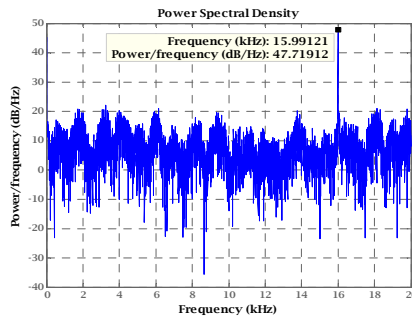


Fig. 4. Spectrum of interference affected DSSS signal

Fig5 depicts the spectrum of DSSS signal after the suppression of interference using LMS algorithm

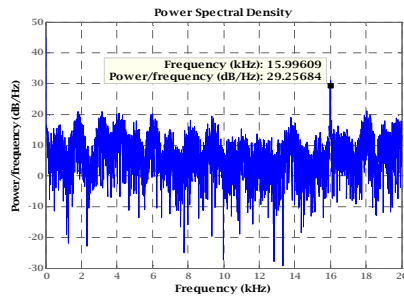


Fig. 5. Spectrum of DSSS signal after interference suppression using LMS algorithm

Fig6 depicts the spectrum of DSSS signal after the suppression of interference using RLS algorithm

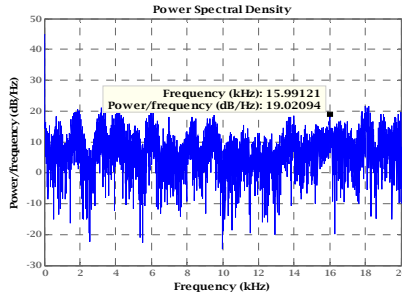


Fig. 6. Spectrum of DSSS signal after interference suppression using RLS algorithm

Fig7 depicts the Bit Error Rate (BER) plot before and after the suppression of interference

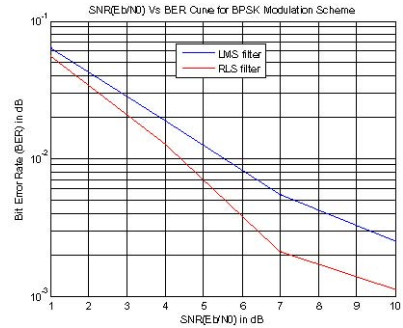
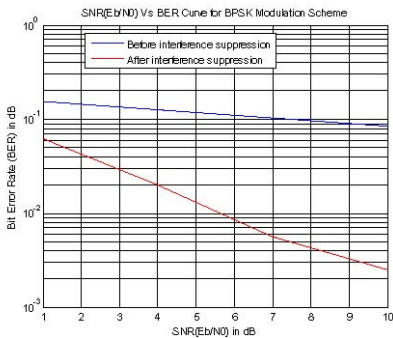


Fig. 7. and 8. Bit Error Rate (BER) plot before and after the suppression of interference and using LMS and RLS filters

6 Conclusion

There is noticeable spectrum reduction near the interference point indicating successful suppression of the interference signal. When compared to LMS filters, RLS filters are more efficient in noisy environment which is evident from fig8. Bit error rate also improved after interference suppression using adaptive filters. Bit error rate improvement is more when the suppression of interference is done using RLS algorithm compared to LMS algorithm which is evident from Fig8.

References

1. Narrow-band Interference Suppression in DSSS Systems Using Efficient Adaptive Filters. In: International Conference on Communications and Mobile Computing (2009)
2. Proakis, J.G., Salehi, M.: Communication Systems Engineering, 2nd edn.
3. Haykin, S., Kailath, T.: Adaptive Filter Theory, 4th edn.
4. Chakrabarti, N.B., Datta, A.K.: Introduction to the principles of digital communication. New Age Publishers (2007)
5. Spread Spectrum Systems- Lecture 18 - EE 359: Wireless Communications – Fall (2009)
6. Lops, M., Tulino, A.M.: Simultaneous Suppression of Multiaccess and Narrow-Band Interference in Asynchronous CDMA Networks. IEEE Transactions on Vehicular Technology 49(5) (September 2000)
7. Wang, C., Ma, M., Ying, R., Yang, Y.: Narrowband Interference Mitigation in DS-UWB Systems. IEEE Signal Processing Letters 17(5) (May 2010)

An Improved Order Estimation of MSF for Stereophonic Acoustic Echo Cancellation

Asutosh Kar¹, Alaka Barik², and Ravinder Nath³

¹ Dept. of Electronics and Telecommunication Engineering, IIT, Bhubaneswar, India

² Dept. of Electronics and Comm. Engineering, SOA University, Bhubaneswar, India

³ Dept. of Electrical Engineering, NIT, Hamirpur, India

asutosh@iiit-bh.ac.in, abarik.nith@gmail.com, nath@nitham.ac.in

Abstract. In this paper the order estimation of each sub-filter, in Multiple Sub-Filters (MSF), parallel structure has been used for Stereophonic Acoustic Echo Cancellation (SAEC). The performance of the MSF with suitable sub-filter order has been studied for two types of algorithms namely; different error algorithm and common error algorithm. The performances of these filter structures and adaptive algorithms have been compared with conventional Single Long Filter (SLF) echo canceller via computer simulations. The order estimation algorithm designed to estimate the order of each sub-filter in MSF not only reduces the computational complexity but also perform better at relatively low SNR ranges. Simulation and results show that MSF with both adaptation algorithms provide better convergence speed as compared to the conventional SAEC.

Keywords: SAEC, MSF, NLMS, filter order, SLF.

1 Introduction

To improve voice quality and have a feel like natural conversation over telephone, two channels audio i.e. stereo system is necessary. Stereophonic devices comprise of two separate acoustic transmission channels that allow listeners to enhance sound realism. Stereo systems consist of a full duplex stereo transmission channel, two loudspeakers and two microphones both in transmission room as well as receiving room. Thus in the stereophonic system the echo problem becomes more complex, and its cancellation is far more difficult to solve than single-channel system, as the input signals for two channels are highly correlated [1]. A variety of methods have been proposed like; adding and modulating little quantities of independent noise to each channel [2], [3], comb filtering [4] and frequency shifting of one channel relative to other [1]. But these methods reduce the quality of stereo sound. Usually long length filter with Least Mean Square (LMS) adaptive algorithm is required for Stereophonic Acoustic Echo Cancellation (SAEC) but suffers from problem of slow convergence. Though Recursive least square (RLS) algorithm can overcome the problem of slow convergence of long length adaptive filter, it is not preferred due to high computational complexity and find difficulty in implementing directly in real time with current DSP processor [5], [6]. In this paper, MSF approach of SAEC is

presented in which the input vector is decomposed into sub-vectors and applied as input to respective sub-filters. The filter length or order is one of the most important parameters of the linear adaptive filter [7], [8], [9]. A too short order filter results in inefficient model of the system and increases the mean square error (MSE) [10], [11], and a too long order filter introduce adaptation noise and extra complexity due to more taps [12]-[15]. Therefore to balance the adaptive filter performance and complexity there should be an optimum order of the filter. Further, In this paper, emphasis has been given on the order estimation of each sub-filter in MSF used for SAEC. The proposed algorithm finds the desired order in a time varying environment with reduced complexity. The modified weight update equation and variable error width in the proposed order estimation algorithm makes it independent of initialization and improved convergence even at high eigen value spread scenarios.

The paper consists of 5 sections. MSF based approach for SAEC has been discussed in section 2 with different error algorithm (DEA) and common error algorithm (CEA). In section 3, the filter order optimization of MSF in SAEC has been presented. The computer simulation setup and results depicting the advantage of using MSF bases SAEC along with the variable order estimation of MSF have been presented in section 4. Conclusion of the paper is discussed in section 5.

2 MSF Approach to SAEC

The length of the acoustic echo path dynamically varies depending on the environment. So the computational complexity of the stereophonic acoustic echo cancellation (SAEC) is very high and critically dependent on the echo cancellation algorithm along with the length of filter used. Using a fixed length long adaptive filter, the adaptive algorithm becomes very slow in terms of convergence speed and it introduces adaptation noise due to mismatch of extra coefficients [15]. Thus to overcome the slow convergence of long length LMS adaptive filter it is split into 2M sub filters, where M number of sub-filter for each channel, for better convergence performance. Total 2M number of individual sub-filter outputs will be generated by

$$y_i(n) = X_{p,k}^T(n)W_i(n) \tag{1}$$

where $i = 0 \dots 2M - 1$ and

$$p = \begin{cases} 1 \text{ and } k = 0 \dots M - 1 \forall i = 0 \dots M - 1 \\ 2 \text{ and } k = 0 \dots M - 1 \forall i = M \dots 2M - 1 \end{cases}$$

First M subfilters outputs $[y_0(n), \dots y_{M-1}(n)]$ will be obtained from the input of channel one as $[X_{1,0}(n), \dots X_{1,M-1}(n)]$. The next M number of outputs $[y_M(n), \dots y_{2M-1}(n)]$ will be obtained from input of channel two with input $[X_{2,0}(n), \dots X_{2,M-1}(n)]$.

2.1 Different Error Algorithm for SAEC

For different error algorithm (DEA), different error signals are used for adaptation as shown in Fig. 1. This algorithm has a higher steady state error than the conventional one. The error signal for different error method can be given as [16]

$$\begin{aligned}
 e_0(n) &= d(n) - \mathbf{X}_{1,0}^T(n) \mathbf{W}_0(n); \\
 e_i(n) &= e_{i-1}(n) - \mathbf{X}_{p,k}^T(n) \mathbf{W}_i(n);
 \end{aligned}
 \tag{2}$$

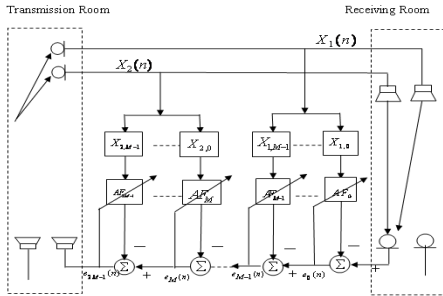


Fig. 1. Adaptation Scheme for MSF based SAEC (Different Error Algorithm)

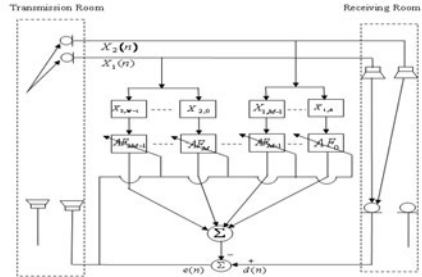


Fig. 2. Adaptation Scheme for MSF based SAEC (Common Error Algorithm)

The LMS adaptation for different error method for SAEC is

$$\begin{aligned}
 \mathbf{W}_0(n+1) &= \mathbf{W}_0(n) + \mu_0 e_0(n) \mathbf{X}_{1,0}(n) \\
 \mathbf{W}_i(n+1) &= \mathbf{W}_i(n) + \mu_i e_i(n) \mathbf{X}_{p,k}(n)
 \end{aligned}
 \tag{3}$$

2.2 Common Error Algorithm for SAEC

In common error algorithm (CEA), a common error is used for weight adaptation. A schematic for this is shown in Fig.2. [16]

The error signal in case of common error can be given as

$$e(n) = d(n) - \mathbf{X}_{p,k}^T(n) \mathbf{W}_i(n)
 \tag{4}$$

where $i = 0 \dots 2M - 1$ and

$$p = \begin{cases} 1 \text{ and } k = 0 \dots M - 1 \forall i = 0 \dots M - 1 \\ 2 \text{ and } k = 0 \dots M - 1 \forall i = M \dots 2M - 1 \end{cases}$$

$$\mathbf{W}_i(n+1) = \mathbf{W}_i(n) + \mu e(n) \mathbf{X}_{p,k}(n);
 \tag{5}$$

where $i = 0, 1 \dots 2M - 1$

In both of these proposed algorithms the MSFs used are of fixed length which makes it inefficient in a time varying environment. If the echo has been created inside car where impulse response is smaller than a big hall then a long length filter unnecessary increases the complexity of overall design and extra adaptation noise due to mismatch of extra coefficients also. Similarly a too short order fixed length filter results in under-modelling [15]. Here a new approach for dynamic order estimation has been proposed based on fractional order estimation methods [16], [17].

3 Order Optimization of MSFs Using Variable Length NLMS

There are three different parts in the proposed variable length LMS algorithm:

- Variable Order/tap-length estimation of Adaptive filter
- Adaptive filter weights update
- Updating Step-size

SAEC setup has been used for simulation purpose. The resulting error signal can be modeled as;

$$e_k^{(P)} \equiv d(n) - W_p(1:K)X_p(1:K) \tag{6}$$

for $1 < K < P(n)$

where W_p and X_p correspond to the steady state weight and input vector respectively with respect to tap-length p . The segmented steady state MSE can be further defined as [17],[18]

$$Q_k^{(P)} = E|e_k^{(P)}|^2 \tag{7}$$

With this the modified cost function for searching the optimum order can be defined as [18]

$$\min \{P|Q^{(P)}_{P-\Delta} - Q^{(P)}_P \leq \delta'\} \tag{8}$$

Neglecting the adaptation noise $Q_{P-\Delta}$ and Q_P corresponds to the MMSE for P and $P - \Delta$ order respectively and $Q^{(P)}_{P-\Delta} - Q^{(P)}_P \geq Q_{P-\Delta} - Q_P$ and the equality holds good when $Q_{P-\Delta} = Q_P$ [17],[19].

In SAEC the filter order is always in terms of thousands. So convergence and steady state error optimization can be achieved by adjusting the error width Δ . The need is faster convergence and a small steady state error. Δ provides the trade-off between the convergence rate and steady state bias. The variable error width $\Delta(n)$ can be obtained as $\Delta(n) = \max_{\Delta_{\min}, \Delta_{\max}} * S$ [17]. The min and max value of the error width is fixed according to [18]. The changing factor S can be calculated both for DEA and CEA separately.

$$S^D_{(DEA)} = \frac{\left| \sqrt{\sum_{i=1}^{2M-1} \hat{e}_i^2(n)} - \sqrt{\sum_{i=1}^{2M-1} \sigma_{v,i}^2} \right|}{\sqrt{\sum_{i=1}^{2M-1} \hat{e}_i^2(n) - \sum_{i=1}^{2M-1} \sigma_{v,i}^2} + \sqrt{\sum_{i=1}^{2M-1} \sigma_{d,i}^2} - \sqrt{\sum_{i=1}^{2M-1} \sigma_{v,i}^2}} \tag{9}$$

where $e_i(n) = \zeta_i(n) + v_i(n)$ for $i = 1, 2, \dots, 2M - 1$. if $v_i(n)$ is assumed to be independent of $\zeta_i(n)$ then

$$E[e_i^2(n)] = E[\zeta_i^2(n)] + E[v_i^2(n)] \tag{10}$$

Where $E[e_i^2(n)]$ is the MSE, $E[\zeta_i^2(n)]$ is the excess MSE (EMSE) and the mean square of system noise is defined as $E[v_i^2(n)]$. The value of EMSE increases to large value at

the early stage and comes down to small value when filter order adaptation takes place; it is being used in updating the smoothing parameter S. Similarly S for CEA can be defined as ;

$$S^C_{(CEA)} = \frac{|\tilde{e}^2(n) - \sigma_v^2|}{|\tilde{e}^2(n) - \sigma_v^2| + \sigma_d^2 - \sigma_v^2} \tag{11}$$

Where $e(n) = \zeta(n) + v(n)$ and $E[e^2(n)] = E[\zeta^2(n)] + E[v^2(n)]$ shows the estimated value for $S^C_{(CEA)}$ calculation.

The variable filter order is obtained by the following adaptation proposed [17],

$$P_{nf}(n) = [P_{nf}(n-1) - K_n] + [(e_{P(n)}^{P(n)}(n))^2 - (e_{P(n)-\Delta(n)}^{P(n)}(n))^2] \bar{K}_n \tag{12}$$

where the factor K_n prevents the order to be increased to a unexpectedly large value and \bar{K}_n can be termed as the step size for filter order adaptation. Both K_n and \bar{K}_n designed based on a leaky factor which depends on the system of application [17].

Finally the tap-length $P(n+1)$ in the adaptation of filter weights for next iteration can be formulated as follows [17], [19]:

$$P(n+1) = \begin{cases} \langle P_{nf}(n) \rangle & \text{if } |P(n) - P_{nf}(n)| > \psi \\ P(n) & \text{otherwise} \end{cases} \tag{13}$$

$P = \langle P_{nf} \rangle$ to best balance the system performance and ψ is a very small positive number. $\langle . \rangle$ is to round the fractional value to the nearest integer ψ in the order adaptation equation can be found with K_n and \bar{K}_n i.e. $\psi = K_n \cdot \bar{K}_n^{-1}$.

In [11], [12] the proposed algorithm is based on LMS and the stability condition need to be checked each time the order changes. So it is advocated to use the normalized LMS (NLMS) algorithm for better convergence and constant level of misadjustment [12]-[15].

$$W(n+1) = W(n) + \frac{\bar{\mu}}{X^T(n)X(n)} X(n)e(n) \tag{14}$$

Here $\bar{\mu}$ is the step size for NLMS algorithm. NLMS converges to mean square for condition [7]

$$0 < \bar{\mu} < 2 \tag{15}$$

In order to accelerate the convergence of the algorithm, the step size of the coefficients in the proposed algorithm is updated according to the variable order,

$$W_{P(n)}(n+1) = W_{P(n)}(n) + \frac{\mu'}{X^T_{P(n)}(n)X_{P(n)}(n)[2+P(n)]} X_{P(n)}(n)e^{P(n)}(n) \tag{16}$$

Where μ' is a constant, $\sigma_x^2 = X^T(n)X(n)$ is the variance of input signal. $P(n)$ is the instantaneous variable adaptive tap-length obtained from the proposed fractional order

estimation algorithm. $W_{P(n)}$ and $X_{P(n)}$ are the weight and input vector pertaining to the order $P(n)$ [17];

Let

$$\mu(n) = \frac{\mu'}{\sigma_x^2 [2 + P(n)]} \quad (17)$$

Then (16) can be written as

$$W_{P(n)}(n+1) = W_{P(n)}(n) + \mu(n) X_{P(n)}(n) e^{(P(n))} \quad (18)$$

which forms a variable step LMS (VLMS) algorithm where the step size depends on the order estimation.

4 Simulation Data and Results

In the simulation, transmission room impulse response g_1 , g_2 are generated using hamming window and fir1. The receiving room impulse responses h_1 , h_2 are generated using Hanning window and Kaiser Window. The source signal $s(n)$ in the transmission room is a sample function of a zero mean unity variance Gaussian random process. The two microphone signals were obtained by convolving $s(n)$ with two transmission room impulse responses g_1 , g_2 . The transmission room signals are obtained as convolving the source signal with the transmission room impulse response which is fed to the receiving room. The microphone signal $d(n)$ in the receiving room is obtained by summing the two convolutions $h_1 * X_1$ and $h_2 * X_2$ where $*$ denotes convolution. All simulations have been carried out on MATLAB 7.9. For evaluation purpose, we have compared the two-channel MSF algorithm with the conventional SLF two-channel LMS algorithm. The length of the filter is taken as 1000 and a white noise sequence of 40dB SNR is added to $h_2 * X_2$ with each MSF carrying 500 weights as M is set to 2 as shown in Fig. 3. Similarly in Fig.4 L has been increased to 1200 with $M=3$ and each MSF carrying 400 taps. Comparison has been made between SLF and MSF approach and in both the cases the MSF approach shows better convergence performance than the SLF.

The total number of MSFs used can be increased and is always an integer value. If the order of MSF is fixed the overall length of adaptive filter is also fixed which makes the system inefficient as the room impulse response varies from inside car environment i.e. a short room impulse response to a conference hall i.e. a long room impulse where SAE has occurred. In a time varying scenario this length can be automatically varied in both the direction in accordance with order estimation algorithm proposed in (12). Simulation has been done under this scenario to find the optimum order of each MSF which best adapts the structure of adaptive filter. The MSE performance with filter order variation has been shown in Fig.5. The ideal choice of length for MSF is made at 500 and 400 taps respectively for the simulations in fig.5. The blind initialization of filter order is made at order 10 and 90 respectively. This is done to show that the proposed algorithm is independent of initialization. The MSE decreases with increase in filter order and remains constant after 250 and 300 taps for 400 and 500 fixed ideal length selection respectively as shown in Fig.5.

The system designed with the proposed taps results in same performance as the filter with fixed order as shown in Fig.6. The overall complexity in design decreases by reducing 150 and 200 less weights per MSF i.e. reducing $(3 \times 150 = 450)$, $(2 \times 200 = 400)$ taps without sacrificing the performance. Further, Fig. 6 shows the issue of undermodelling with less number of weights than the required and the better convergence performance with order estimated by the proposed algorithm than fixed length adaptive filter in a dynamic environment.

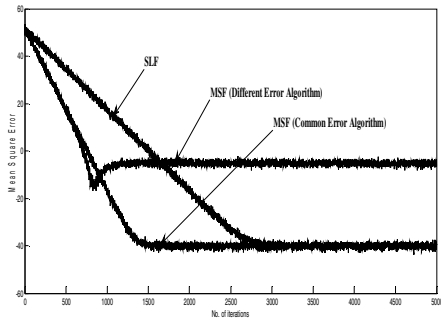


Fig. 3. Comparison of MSF and SLF for SAEC at $L=1000, M=2, SNR=40dB$

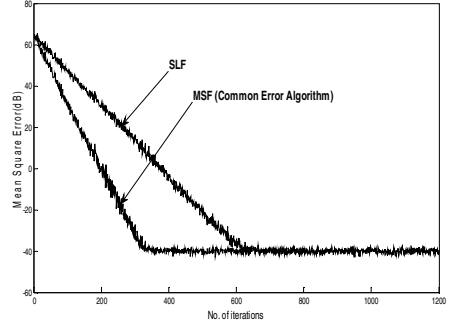


Fig. 4. Comparison of MSF and SLF for $L=1200 M=3$ and $SNR=40dB$

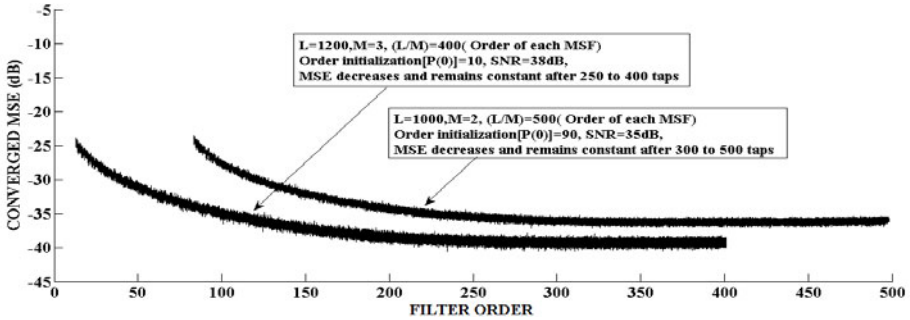


Fig. 5. Converged MSE Vs Filter Order

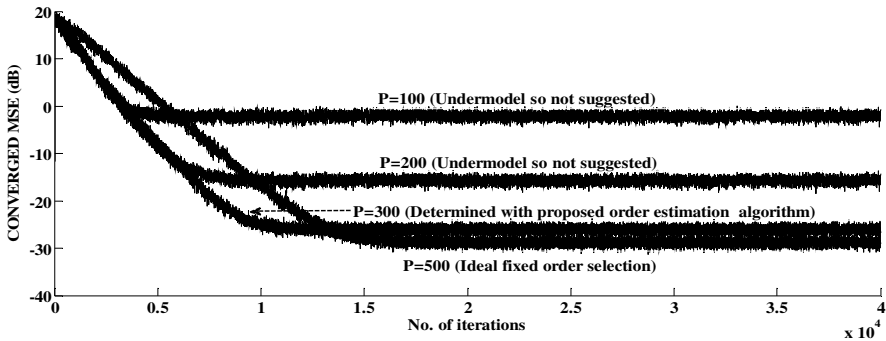


Fig. 6. Converged MSE Vs No. of iterations

5 Conclusion

In this paper, an SAEC based on MSF has been studied via computer simulation and compared to that of SLF. It has been found that MSF outperforms SLF in terms of convergence speed. MSF structure has been studied for two types of algorithms. The DEA gives better convergence speed but at the cost of steady state error while the CEA overcomes the high steady state error problem but with a sacrifice in convergence speed. A modified variable order NLMS based algorithm was proposed that improves the convergence and finds the optimum order of the MSFs in a dynamic environment.

References

1. Benesty, J., Morgan, D.R., Sondhi, M.M.: A Better Understanding and an Improved Solution to the Specific Problems of Stereophonic Acoustic Echo Cancellation. *IEEE Trans. Speech Audio Process.* 6, 156–165 (1998)
2. Sondhi, M.M., Morgan, D.R.: Acoustic Echo Cancellation for Stereophonic Teleconferencing. In: *IEEE Workshop on Appls. of Signal Processing Audio Acoustics*, pp. 141–142 (October 1991)
3. Shimauchi, S., Makino, S.: Stereo Projection Echo Canceller With True Echo Path Estimation. In: *IEEE International Conference on Acoustics Speech and Signal Process.*, vol. 5, pp. 3059–3062 (May 1995)
4. Benesty, J., Morgan, D.R., Hall, J.L., Mohan Sondhi, M.: Stereophonic Acoustic Echo Cancellation Using Nonlinear Transformation And Comb Filtering. In: *IEEE International Conference on Acoustics, Speech and Signal Process.*, vol. 6, pp. 3673–3676 (May 1998)
5. Eneroth, P., Gay, S.L., Gaensler, T., Benesty, J.: An Implementation of a Stereophonic Acoustic Echo Canceller on a General Purpose DSP. In: *Proc. ICSPAT* (1999)
6. Benesty, J., Amand, F., Gilloire, A., Grenier, Y.: Adaptive Filtering Algorithms For Stereophonic Acoustic Echo Cancellation. In: *Proc. ICASSP*, pp. 3099–3102 (1995)
7. Widrow, B., Sterns, S.D.: *Adaptive Signal Processing*. Prentice Hall Inc., Englewood Cliffs (1985)
8. Haykin, S., Haykin, S.: *Adaptive Filter Theory*. Prentice Hall Inc., Englewood Cliffs (1996)
9. Shnyk, J.J.: Frequency Domain and Multirate Adaptive Filtering. *IEEE Signal Processing Magazine* 9(1), 14–37 (1992)
10. Gu, Y., Tang, K., Cui, H., Du, W.: Convergence analysis of a deficient-length LMS filter and optimal-length to model exponential decay impulse response. *IEEE Signal Process. Lett.* 10, 4–7 (2003)
11. Mayyas, K.: Performance analysis of the deficient length LMS adaptive algorithm. *IEEE Trans. Signal Process.* 53(8), 2727–2734 (2005)
12. Gong, Y., Cowan, C.F.N.: A novel variable tap-length algorithm for linear adaptive filterers. In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Montreal, QC, Canada (May 2004)
13. Gong, Y., Cowan, C.F.N.: An LMS style variable tap-length algorithm for structure adaptation. *IEEE Trans. Signal Process.* 53(7), 2400–2407 (2005)
14. Gong, Y., Cowan, C.F.N.: Structure adaptation of linear MMSE adaptive filters. *Proc. Inst. Elect. Eng., Vis., Image, Signal Process.* 151(4), 271–277 (2004)

15. Schüldt, C., Lindstromb, F., Li, H., Claesson, I.: Adaptive filter length selection for acoustic echo cancellation. *Signal Processing* 89, 1185–1194 (2009)
16. Barik, A., Murmu, G., Bhardwaj, T.P., Nath, R.: LMS adaptive multiple sub-filters based acoustic echo cancellation. In: *Proc. 2010 IEEE Int. Conf. Computer and Communication Technology (ICCT 2010)*, Allahabad, India, pp. 824–827 (September 2010)
17. Kar, A., Nath, R., Barik, A.: A VLMS based pseudo-fractional order estimation algorithm. In: *ACM Sponsored International Conference on Communication, Computing and Security (ICSSS 2011)*, Rourkela, India, pp. 119–123 (February 2011)
18. Li, N., Zhang, Y., Hao, Y., Zhao, Y.: A new variable tap-length LMS algorithm with variable error width. In: *Proc. IEEE Int. Conf. on Signal Processing*, pp. 276–279 (October 2008)
19. Kar, A., Nath Sharma, R.: A unique automatic and dynamic order estimation for linear adaptive filters. In: *IEEE Sponsored International Conference on Intelligent System and Data Processing (ICISD 2011)*, Gujarat, India (January 2011)

Application of Stochastic Model on Routing Technique in Multi Class Queueing Network

K. Sivaselvan¹ and C. Vijayalakshmi Seshathri²

¹ Department of Mathematics, Jeppiaar Engineering College, Chennai

² Department of Mathematics, Sathyabama University, Chennai
sivajpr@gmail.com, vijusesha2002@yahoo.co.in

Abstract. In communication networks, the network size is growing hasty and the computation effort to find a path between the source–destination pairs is increased massively. Multiple paths may exist between the source-destination nodes which direct that traffic load variations, overhead, response time take place. Routing plays a vital role on the performance and functionality of computer networks. Routing networks means identifying a path in the network that optimizes a certain criterion which is called as Quality of Service (QoS) routing and it is failure in the environment of large scale networks. The storage and updating cost of routing procedure is prohibitive as the number of nodes in the network gets large. Stochastic techniques have assumed a prominent role in computer graphics, because of their success in modeling a variety of complex and natural phenomena. The usefulness of a particular stochastic model depends on both its computational advantages and on the extent to which can be adjusted to describe different phenomena. Network isolation is a key solution for improving the scalability problem in large networks. The main aim of isolation is minimizing the computation effort by maximizing the probability of having a path between source-destination pairs in the network. This paper deals with the specification and analysis of routing procedures that are effective for large hoard and promote packet switched computer networks. The new concept of stochastic isolation method introduced to resolve the scalability in Quality of Service routing algorithm. Graphical representation shows that how the new method improves the performance measure in terms of reduction in computational effort.

1 Introduction

Broadband of integrated service network is expected to support applications with Quality of Service requirements. The design of data network in integrated service is highly depends on the source-destination pair, which dispute for simplicity in the network central part and intricacy at the end hosts. This design principle enabled that the optimization performance within the network by computational policy. A communication network consists of a set of nodes that are connected by a set of links. A path has defined in the network where a collection of sequential communication links eventually connecting two nodes to each other. The process of finding and selecting the paths in the network is termed as routing function. A routing policy is a decision

rule that selects which nodes to take next based on the current time and realize network link. The objective of routing technique is (i) distribute and searching the state information in optimal way of the network (ii) how to reduce the computational effort in searching for a path. The main drawback of all modern routing algorithms is in lack of ability to scale the large networks proficiently. In traffic pattern, the router allows to share the network congestion state previously examined by other connections sharing the same bottleneck link, which improves the throughput drastically. For such requirements, transition is to isolate the central part router function from the computational framework. The stochastic isolation method that dynamically changes network isolation according to traffic patterns in the network in order to minimize an objective function that reflects the computational effort involved in routing algorithm used in the network. Network isolation is the solution to enhancing the scalability in large networks. Network isolation decomposes a network into sub networks according to particular rules and considerably reduces the computational effort of routing. In this method, the probabilities used to partition the network correspond to the frequency of connection requests between every pair of nodes in the network. The stochastic isolation technique has reduces the computational effort in routing network and showed that which is effective and efficient for large scale routing network. The rest of the paper is organized as follows: Section 3 explain the concept of QoS route sharing resource. Section 4 describes the routing computational framework of packet switching network. Section 5 discusses stochastic isolation and introducing some notations. Section 6 explains the overhead in scalability techniques. Section 7 graphical representation discussed. Finally, section 8 concludes the paper.

2 Literature Survey

Broadband of integrated Scalability in communication network had been developed by Amitabh Mishra(2002). Ariel Orda et al (2002), has clearly explained a scalable approach to the partition of QoS requirements in unicast and multicast. Fang Hao et al (2002), had explained the scalable QoS routing performance evaluation of topology aggregation. P. Gupta et al (2006), has clearly envisaged the optimal throughput allocation in general access networks. W. Ching et al (2009), has analyzed the optimal service capacities in a competitive multiple-server queueing environment. IE. Leonardi et al (2005), had approached the Joint optimal scheduling and routing for maximum network throughput. X. Lin et al (2004), had analyzed an optimization based approach for quality of service routing in high-bandwidth networks. Orda et al (2003), had approached the pre-computation schemes for QoS routing. S. Sinha Deb et al (2003), had given a detailed explanation of a new approach to scale quality of service routing algorithms.

3 Quality of Service on Routing Update Resources

The role of a QoS routing policy is to compute a suitable path for the different types of traffic generated by the various applications, while maximizing the utilization of network resources. The implementation of these objectives requires the development

of algorithms that find multi-constrained paths taking into consideration the state of the network and the traffic requirements namely, delay jitter, loss rate and available bandwidth. Routing decision is determined based on the network information available at the source. A packet flow has referred to as IP packet stream linked from a source to a destination with QoS in finite period. When a flow request arrives, there is routing algorithms that stipulate the routes to all known destination, which preserve the position of all the routes in the network. Recurring state is the main part in routing algorithms which updating the minimum distance of each router for all destinations. The nodes have to be update periodically by routing update before the time run out. For the services that have ensured the Qos guarantee for the entire period of flow, otherwise the flow request is starve. The main aim of QoS routing is to minimize the run time and has to reduce the impact on the run set-up time at least from the routing computation point of view. Consider a packet switching network with a source-destination pair $[i, j]$. Let $P_{[i,j]}(\eta)$ denotes the set of paths determined at time η and $Z^{\kappa}_{[i,j]}(\eta)$ signifies the set of links where $\kappa \in P_{[i,j]}(\eta)$. The average time of the path κ for $[i, j]$ is determined by $\Psi_{[i,j]}(t) = \min_{b \in Z^{\kappa}_{[i,j]}} \{X_b(t)\}$; $X_b(t)$ represents the available capacity of link b at time t .

4 Routing Computational Framework

The basic component of Quality of Service routing framework is path selecting that can operate in a link state routing protocol environment where different information can be used in two different scales. The goal of routing computational framework is (i) to reduce the impact of flow setup time. (ii) to avoid user level re-attempt in heavily loaded network (iii) to select a route quickly in possible paths. The framework consists of three stages at different time scale: (i) First Round Path Communicating (FRPC) stage. (ii) Sorted Path Ordering (SPO) stage (iii) Definite Route Assortment (DRA) stage.

The First Round Path Communicating (FRPC) stage does preliminary determination of a set of possible paths from a source node to destination node. The Sorted Path Ordering (SPO) stage follows Markov process (selects the most recent states of all links available to each node) and filters it to provide a set of Quality of Service acceptable paths. Moreover, this phase order the routes from most to least acceptable paths which is obtained from list of FRPC stage. The Definite Route Assortment (DRA) stage follows that to select a definite route as swiftly as possible based on the pruned available paths from the SPO stage. The main advantage of this framework is that various distributed routing schemes can fit into this framework and multiple Quality of Service requirements be used.

4.1 Packet Routing Policy

In satellite, the communication sub network accomplishes communication among the network resources. In packet switching network, the message has wrecked into small segments and then transmitted through the network in the form of hoard and promote

switching. A packet has transmitted from source node to destination node, which may be hoard in queue at any intermediate node for transmission and then promote to the next node. The selection of the next node is based on the routing policy. Routing policy has divided into two categories: Deterministic (Design phase) and Adaptive (Networks Operation). Adaptive policy plays a vital role for triumphant operation of networks and it describes the state of the network. A central node providing the routing information to all sub nodes in the network which computing the information directly.

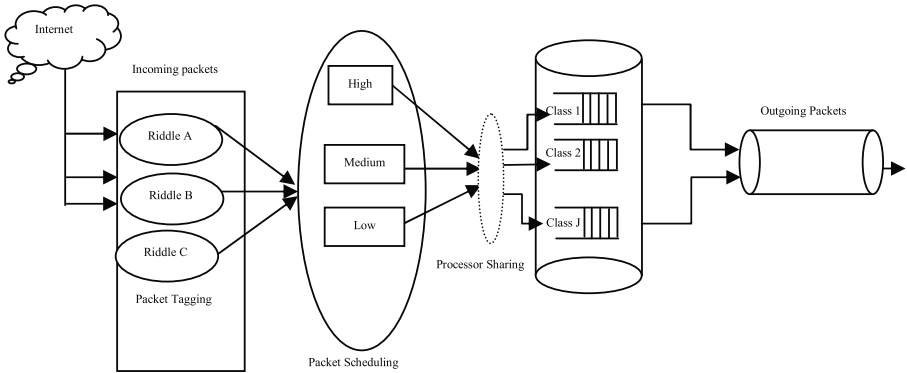


Fig. 1. Queueing Discipline in Packet Switching Network

5 Stochastic Isolation

Scalability Stochastic isolation technique is designed to partition the original network into a number of blocks which enhancing the scalability routing algorithm in large networks. In this method network is partition in a probabilistic manner that corresponds to the frequency of connection requests between every pair of nodes in the network. The main objective of stochastic isolation is to minimize the mean computational effort spent by the routing algorithms used in the network by maximizing the chance of selecting a source–destination pair in the same block of the partition. In particular, if the source-destination pair in the same block there must exist at least one path between every pair of nodes in the block, which is termed as irreducible. For low connectivity network, the partition is more difficult with irreducible blocks whereas easily constructed in high connectivity network. Thus, each block should consist of at least two nodes in order to have communication between them.

The component of Quality of Service routing framework is path selecting that can operate in a link state routing protocol environment where different information can be used in two different scales. The goal of routing computational framework is (i) to reduce the impact of flow setup time. (ii) to avoid user level re-attempt in heavily loaded network (iii) to select a route quickly in possible paths. The framework consists of three stages at different time scale: (i) First Round Path Communicating (FRPC) stage. (ii) Sorted Path Ordering (SPO) stage (iii) Definite Route Assortment (DRA) stage.

The First Round Path Communicating (FRPC) stage does preliminary determination of a set of possible paths from a source node to destination node. The Sorted Path Ordering (SPO) stage follows Markov process (selects the most recent states of all links available to each node) and filters it to provide a set of Quality of Service acceptable paths. Moreover, this phase order the routes from most to least acceptable paths which is obtained from list of FRPC stage. The Definite Route Assortment (DRA) stage follows that to select a definite route as swiftly as possible based on the pruned available paths from the SPO stage. The main advantage of this framework is that various distributed routing schemes can fit into this framework and multiple Quality of Service requirements be used.

5.1 Objective Function of Isolation Strategy

The objective is to minimize a quantitative measure for any network isolation structure in terms of the computational effort in routing. The mean computational effort in finding a path satisfy the constraints from a source node to destination node, averaged overall source-destination pairs, in a network of K nodes and Z links be $\Psi(K, Z)$. The objective function of computational effort is defined as

$$C(K, Z) = \min \left[\bigcup_{i=1}^N P_i \Psi_i(K_i, Z_i) + (1 - \bigcup_{i=1}^N P_i) \Psi(K, Z) + P(K, Z) \right]$$

where P_i represents the probability that given a connection request when both source and destination nodes located in the same block, $P(K, Z)$ represents the isolation overhead per connection request, $\bigcup_{i=1}^N P_i \Psi_i(K_i, Z_i)$ represents the computational effort involved when both source and destination nodes located in the same block and $\left[1 - \bigcup_{i=1}^N P_i \right] \Psi(K, Z)$ represents the computational effort when both source and destination nodes located in the different blocks.

Let P_{ij} be the conditional probability that given a connection request from source node 'i' to the destination node 'j'. P_{ij} is defined as follows:

$$P_{ij}(\omega) = \frac{\eta_{ij}(\omega)}{\sum_{\substack{i, j \in N \\ i \neq j}} \eta_{ij}(\omega)}$$

where $\eta_{ij}(\omega)$ denotes the number of times source node i has requested a connection to destination node in the last ω time unit.

$P_{ij}(\omega)$ will be more accurate, as the time unit ω increases $\lim_{\omega \rightarrow \infty} P_{ij}(\omega) = P_{ij}$

$$P_{ij}(b) = \frac{C_{n_b-2}^{N-2}}{C_{n_b}^N} = \frac{(b-2)! (N-2-b+2)!}{(N)!} = \frac{(n_b-1)(n_b)}{(N-1)N}$$

where n_b denotes the number of nodes in block b and N denotes the number of nodes of the network.

6 Scalability Techniques Overhead

The most important constraint on isolation is partition overhead. There are two major types of overhead are routing update reduction and route computation reduction. The routing update reduction provides the information continuously updated to the network nodes through routing information. Additionally frequent updated routing leads to a better routing performance in the network and consumes more network bandwidth and processing power. Reduction of routing update frequency in two ways:(1) searching for appropriate routing update trigger policies to provide controllable update frequency and predictable accuracy, (2) designing appropriate routing algorithms to minimize the impact of stale routing information.

The route computation reduction is essential for achieving high-quality routing performance and scalability. Route pre-computation and path catching are the two major approaches in order to reduce route computation. Route pre-computation is used to compute and store the paths to all destinations before the request that leads that minimize the request operations. Moreover, that it helps to compute multiple paths to the same destination nodes and balance the traffic load. Path catching avoids computing the same path again.

7 Graphical Representation

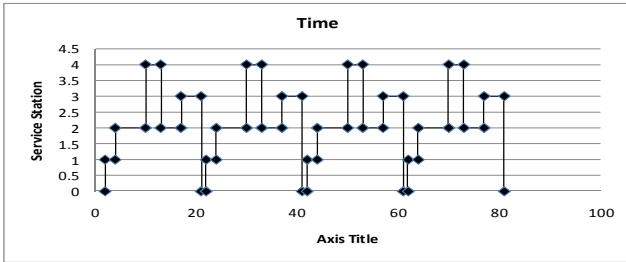


Fig. 2. Service Load Distribution

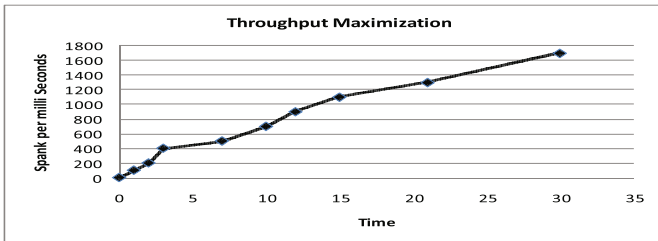


Fig. 3. CPU Utilization

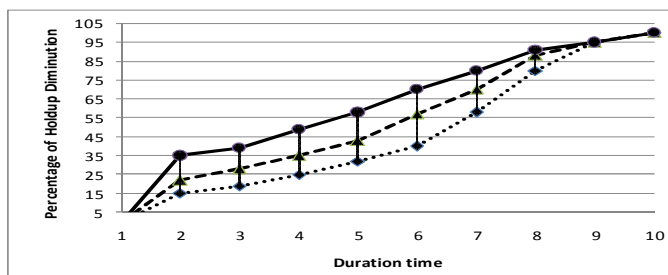


Fig. 4. Dispersion Factor

8 Conclusion

In this paper, The Quality of Service routing is a main component of a computational framework. The scalability network is the challenging issue in the environment of large networks. In this paper, a new concept has introduced for isolation using stochastic isolation to reduce computational effort to find a path. The stochastic isolation technique is to maximizing the scalability and minimizing the complexity in large networks. Graphical representation shows that the stochastic isolation speed up the routing functions.

References

- Amitabh Mishra, I.: Scalability in communication network. *IEEE Networks* 16(4), 10 (2002)
- Orda, A., Sprintson, A.: Precomputation schemes for QoS routing. *IEEE/ACM Trans. Netw.* 11(4), 578–591 (2003)
- Orda, A., Sprintson, A.: A scalable approach to the partition of QoS requirements in Unicast and Multicast. In: *IEEE INFOCOM*, vol. (1), pp. 685–694 (2002)
- Leonardi, E., Mellia, M., Ajmone Marsan, M., Neri, F.: Joint optimal scheduling and routing for maximum network throughput. In: *Proc. IEEE INFOCOM 2005*, Miami, FL, pp. 819–830 (June 2005)
- Hao, F., Zegura, E.W.: On Scalable QoS Routing Performance Evaluation of Topology Aggregation. In: *IEEE INFOCOM 2002*, vol. (1), pp. 147–156 (March 2002)
- Bettahar, H., Bouabdallah, A.: A New approach for Delay-Constrained routing. *Elsevier Publication-Computer Communication* 25, 1751–1764 (2002)
- Zhang, K.A.S., Kelly, T., Stewart, C.: Operational analysis of processor Speed scaling. In: *SPAA* (June 2008)
- Siva Selvan, K., Vijayalakshmi, C.: Algorithmic Approach For the Design Markovian Queueing Network with Multiple Closed Chains. In: *International Conference on TRENDZ Information Sciences and Computing*. IEEE xplore, Sathyabama University TISC (2010)
- Younis, O., Fahmy, S.: Constraint-based routing in the internet: basic principle and recent research. *IEEE Communication Society Surveys & Tutorials* 5, Xg3, 42–56 (2003)
- Gupta, P., Stolyar, A.L.: Optimal throughput allocation in general random access networks. In: *Proceedings of 40th Annual Conf. Inf. Sci. Systems*, pp. 1254–1259 (2006)
- Halabi, S., McPherson, D.: *Internet routing architectures*, 2nd edn. Cisco Press (2000)

- Mao, S., Panwar, S.S., Hou, Y.T.: On minimizing end-to-end delay with optimal traffic partitioning. *IEEE Transactions on Vehicular Technology* 55(2), 681–690 (2006)
- Bhatti, S.N., Crowcroft, J.: QoS-sensitive flows: Issues in IP packet handling. *IEEE Internet Comput.* 4, 48–57 (2000)
- Sinha Deb, S., Woodward, M.E.: A New Approach to Scale Quality of Service Routing Algorithms. In: *Globecom* (2004)
- Korkmaz, T., Krunz, M.: Multi-Constrained Optimal Path Selection. In: *Proceedings of the IEEE INFOCOM*, pp. 834–843 (2001)
- Ching, W.-K., Choi, S.-M., Huang, M.: Optimal Service Capacities in a Competitive Multiple-Server Queuing Environment. In: Zhou, J. (ed.) *Complex 2009*. LNICST, vol. 4, pp. 66–77. Springer, Heidelberg (2009)
- Liu, W., Lou, W., Fang, Y.: An efficient quality of service routing algorithm for delay-sensitive applications. *Elsevier Publication-Computer Networks* 47, 87–104 (2005)
- Lin, X., Shroff, N.B.: An optimization based approach for quality of service routing in high-bandwidth networks. In: Presented at the *IEEE INFOCOM*, Hong Kong, China (March 2004)

An Efficient Data Structure for Document Clustering Using K-Means Algorithm

Ramanji Killani¹, Suresh Chandra Satapathy², and A.M. Sowjanya³

¹ MVGR College of Engineering, Vijayanagaram, India

² ANITS, Vishakapatnam, India

³ A.U. College of Engineering, Andhra University, Visakhapatnam, India

Abstract. In this paper, we proposed an efficient data structure called “Sparse Matrices” for representing documents. The document database can be represented by using sparse matrices rather than dense matrices. The matrix can be given as an input for k-means algorithm. Using sparse matrices not only will reduce the size of the database as well as it found efficient in running the program. The experimental results have shown that sparse matrices gives good results compared to dense matrices.

Keywords: Document clustering, sparse matrices and k-means algorithms.

1 Introduction

Text mining is playing crucial role in mining textual databases. Text Mining is a multidisciplinary field, involving information retrieval, text analysis, information extraction, clustering, categorization, visualization, database technology, machine learning and data mining [1],[2]. Text mining data is useful to indicate similarities at the textual level, but is also affected by the ambiguities of vocabularies [9]. The applications of text mining are document clustering, document classification etc. Document clustering techniques have been receiving more and more attentions as a fundamental and enabling tool for efficient organization, navigation, retrieval and summarization of huge volumes of text documents with a good document clustering method; computers can automatically organize a document corpus into a meaningful cluster hierarchy, which enables an efficient browsing and navigation of the corpus [3]. Document clustering is basically a data clustering procedure specialized to the text document materials. The intention of the text document clustering is to divide the given documents into a certain number of groups while maximizing the similarities between the documents into internal groups and minimizing the similarities between the documents disseminated among the different groups [4]. The process of text document clustering can be separated into three parts: document representation, smoothing data in document and clustering. The purpose of document representation is to convert the document data into numerical format and this can be done by using vector space model. Smoothing is used to reduce the size of the data in the document and can be done by using dimensionality reduction techniques. Clustering can be done by using two techniques i.e, partitional clustering and hierarchical clustering, the search space can be partitioned into clusters where as in hierarchical, the space can be

divided into hierarchical forms. In this work, we have used k-means algorithm for clustering.

Generally, the dense matrix occupies lots of space and k-means algorithm is taking lots of time to give the result. The sparse matrices will take less space and hence the dataset will be in compressed form. So, the mining process with k-means algorithm will run faster.

In this work, we are using document database of huge size and the data is not dense as well. Due to that reason, the database is represented using sparse matrices. The advantage of using sparse matrices is by nature the data is easily compressed, so the space can be reduced and as well as the algorithm runs faster.

The rest of the paper is organized as follows: Section-2 gives basic idea of document representation. In section-3, the sparse matrices are discussed. Clustering algorithm is presented in section-4. The simulation results are given in section-5. Finally the paper is concluded in section-6.

2 Document Representation

2.1 Data Representation for Document Clustering

To find the relevance in the documents, the dataset is represented using vector space model. Each document is represented as a vector and the document database is treated as a vector space. The documents can be represented into vectors in two phases. In the first phase the documents are scanned. This phase identifies the unique words in the document dataset and gives unique term number for each term in the document. In the second phase, the documents are represented in vectors, i.e. with their dimension and the quantity of the dimension. A document d is represented as follows, $d = \{t_1, t_2, t_3 \dots t_m\}$ where $t_1, t_2 \dots t_m$ are terms.

$$d = a_1 t_1 + a_2 t_2 + \dots + a_n t_n \quad (1)$$

Where a_1, a_2, \dots, a_n are frequencies of terms i.e, how many times terms t_1, t_2, \dots, t_n are repeated in the document.

2.2 Picking Important Terms in Documents

Important terms are to be identified before applying clustering techniques. To identify important terms a measure called term-frequency and inverse document frequency (Tf-Idf) is used. The weight of term ' x ' in a document is given by the following equation (2):

$$w_{yx} = tf_{yx} * idf_{yx} \quad (2)$$

2.3 Finding Similar Documents

There are number of methods to calculate similarity between two documents such as Cosine methods, Euclidian method, Minkowski distance measure etc. In our work we have used Minkowski distance measure as it is the generalization of both Euclidian and Manhattan distance.

$$d(m_i, m_j) = \sqrt{\frac{\sum_{k=1}^{dm} (m_{ik} - m_{jk})^2}{dm}} \quad (3)$$

Where m_i, m_j are any two document vectors, dm is the dimension of vector space and m_{ik}, m_{jk} are weights for dimension 'k'.

3 Sparse Matrices

The tabular data can be represented in matrices format in computer's memory. There are two kinds of matrices, sparse matrices and dense matrices. The sparse matrices consists of many elements as zeros where as in dense matrices consists of many elements as non zeros. The examples sparse matrices are Identity matrix. A matrix having only a small percentage of nonzero elements is said to be sparse. In a practical sense an $n \times n$ matrix is classified as sparse, if it has order of n nonzero elements; say two to ten nonzero elements in each row, for large n . The matrices associated with a large class of man-made systems are sparse. For example, the matrix representing the communication paths of the employees in a large organization is sparse, provided that the i th row and the j th column element of the matrix is nonzero if and only if employees i and j interact.

Sparse matrices appear in Information retrieval, linear programming, structural analyses, network theory and power distribution systems, numerical solution of differential equations, graph theory, genetic theory, social and behavioral sciences and computer programming. It is difficult to solve the problems which involves with large matrices which either are impossible to invert on available computer storage or are very expensive to invert. Since such matrices are generally sparse it is useful to know the techniques currently available for dealing with sparse matrices. This allows one to choose the best technique for the type of sparse matrix encounters. The time and effort required to develop the various techniques for handling sparse matrices is especially justified when several matrices having the same zero-nonzero structures but differing numerical values have to be handled [8]. Dense matrices are unsuitable for representing high dimensional data, there is a need for a collection of matrices from real applications where data can be represented by using sparse matrices [7]. Generally, the data in sparse matrices can be represented as row, column and value of the element.

Sparse matrix formats compress large matrices with a small number of nonzero elements into a more compact representation. The goal is to both reduce memory footprint and increase efficiency of operations such as sparse matrix-vector multiplication (SpMV) [10].

For example, let's take I4 matrix. The dense matrix representation for I4 is as follows:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

The sparse matrix representation for I4 is given below:

$$\begin{bmatrix} 0 & 0 & 1 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \\ 3 & 3 & 1 \end{bmatrix}$$

3.1 Memory Comparison between Sparse Matrices and Dense Matrices

The dense matrix occupies huge amount memory especially, in the application of document mining. Let us consider a 500 x 500 matrix with 1994 nonzero elements. The dense matrix will require 500 x 500 x 4 = 1million bytes where as the sparse matrix will takes 3 x 1994 x 4= 23,928 bytes.

3.2 Representation of Sparse Matrix

Here the entire document is represented in sparse matrix vector where each document is a sparse matrix. Let us assume, the document database consists of (d1, d2, ..., dn) documents. Each document is represented as follows:

$$d1 = \begin{bmatrix} col1 & col2.. & coln \\ val1 & val2.. & valn \end{bmatrix} \quad (4)$$

Where col1, col2, ... and coln are column values and row value is 1 and the element values are val1, val2, ..., valn. Here, only non zero values will be stored.

4 Clustering Algorithm

Clustering algorithms are classified into partitional clustering and hierarchical clustering. In this work, we have used K-means Algorithm for document clustering.

4.1 K-Means Algorithm

K-Means algorithm [5] is a partitional-based clustering algorithm. It searches for the solution in local space area. K-Means algorithm divides the problem space area into partitions and searches for the solution. The algorithm takes the input in sparse matrix format. The similarity matrix consists of Tf-Idf values, number of clusters 'n' and initial centroids. The output is in terms of Set of clusters, Intra-cluster distance, Inter-cluster distance. K-means has some disadvantages like the solution is completely depends upon the initial cluster centroids which are generated in random manner and it searches for the solution in local space area and gets trapped in local optima often.

The K-means algorithm can be summarized as follows:

- (1) Randomly select cluster centroid vectors to set an initial dataset partition.
- (2) Assign each document vector to the closest cluster centroids.

(3) Recalculate the cluster centroid vector \mathbf{c}_j using equation 5.

$$C_j = \frac{1}{n_j} \sum_{\forall d_j \in S_j} d_j \quad (5)$$

where \mathbf{d}_j denotes the document vectors that belong to cluster \mathbf{S}_j ; \mathbf{c}_j stands for the centroid vector; \mathbf{n}_j is the number of document vectors that belong to cluster \mathbf{S}_j .

(4) Repeat step 2 and 3 until the convergence is achieved.

5 Experimental Setup and Results

This experiment is performed by using J2SE 6.0. The clusters are developed with K-means algorithm. The three datasets are taken from Tech TC-300 repository [6]. The descriptions are given below:

1. Exp_240218_474717 consists of 185 documents and 6560 terms: data1
2. Exp_22294_25575 consists of 127 documents and 12812 terms: data2
3. Exp_20673_269078 consists of 147 documents and 14600 terms: data3

The numbers of clusters are chosen by user based on the knowledge derived from the contents of the dataset under investigation. For all experiments we have chosen number of clusters ranging from 3 to 7. The good cluster should have minimum intra cluster and maximum inter cluster distances.

The results shown below are of document clustering implemented with dense matrix.

Table 1. Results of Documenter clustering using Dense Matrix

S. No	Dataset Size		Number of Clusters	Space required	Average Intra Cluster Distance	Average Inter Cluster Distance	Time (in milli seconds)
	Number of Documents	Size of dimension					
1	185	6560	3	1.16MB	9.374	13.584	6831.2
2	185	6560	4	1.16MB	8.868	18.341	6945.5
3	185	6560	5	1.16MB	7.309	14.164	6739.4
4	185	6560	6	1.16MB	9.635	18.513	7254.6
5	185	6560	7	1.16MB	6.856	15.640	7139.2
6	127	12812	3	6.21MB	7.441	14.456	8247.5
7	127	12812	4	6.21MB	8.865	16.774	9000.0
8	127	12812	5	6.21MB	6.911	14.449	8894.7
9	127	12812	6	6.21MB	7.207	16.446	9575.0
10	127	12812	7	6.21MB	6.392	14.637	9569.3
11	147	14600	3	8.19MB	8.787	21.162	10650.5
12	147	14600	4	8.19MB	8.158	15.780	10529.7
13	147	14600	5	8.19MB	5.062	13.452	10887.3
14	147	14600	6	8.19MB	5.787	15.848	11686.2
15	147	14600	7	8.19MB	5.883	16.305	11396.8

The results shown below are of document clustering implemented with sparse matrix.

Table 2. Results of Documenter clustering using Sparse Matrix

S. No	Dataset Size		Number of Clusters	Space required	Space occupied	Average Intra Cluster Distance	Average Inter Cluster Distance	Time (in milli second)
	Number of Documents	Size of dimension						
1	185	6560	3	1.16MB	0.35MB	8.32164	12.34232	7678.8
2	185	6560	4	1.16MB	0.35MB	9.65678	16.03120	7137.8
3	185	6560	5	1.16MB	0.35MB	7.30900	14.16478	7087.4
4	185	6560	6	1.16MB	0.35MB	7.67342	19.23376	7378.2
5	185	6560	7	1.16MB	0.35MB	8.4267	15.64021	7099.8
6	127	12812	3	6.21MB	0.24MB	8.12150	12.45698	8578.0
7	127	12812	4	6.21MB	0.24MB	7.86576	16.77452	9034.4
8	127	12812	5	6.21MB	0.24MB	5.39230	14.13302	9109.8
9	127	12812	6	6.21MB	0.24MB	6.86863	13.09628	9393.6
10	127	12812	7	6.21MB	0.24MB	7.42410	12.45108	9359.2
11	147	14600	3	8.19MB	0.35MB	9.21124	16.57858	10880.8
12	147	14600	4	8.19MB	0.35MB	7.34905	15.64275	10768.8
13	147	14600	5	8.19MB	0.35MB	4.83864	13.20175	11100.0
14	147	14600	6	8.19MB	0.35MB	5.28120	14.67351	11453.2
15	147	14600	7	8.19MB	0.35MB	5.21360	12.32568	11694.0

6 Conclusion

By using the sparse matrix, the documents data can be represented. Sparse matrix representation occupies only less memory to store the entire document description. The results are showing that the dense matrix occupies lot of memory compared to sparse matrix. Therefore, the sparse matrix is the good data structure to represent the high dimensional data.

References

1. Tan, A.-H.: Text Mining state of art and challenges. In: Proceedings of the PAKDD 1999 Workshop (1999)
2. Han, J., Kamber, M.: DataMining concepts and Techniques, 2nd edn. Morgan Kaufmann publishers (2006)
3. van der Maaten, L.J.P., Postma, E.O., van den Herik, H.J.: Dimensionality Reduction a Comparative Review. Citeseer (2007)
4. Cui, X., Potok, T.E., Palathingal, P.: Document Clustering using particle swarm optimization. In: IEEE Swarm Intelligence Symposium (2005)
5. MacQueen, J.B.: Some Methods for classification and Analysis of Multivariate Observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297. University of California Press (1967)

6. The TechTC-300 Test Collection for Text Categorization Version: 1.0 TechTC - Technion Repository of Text Categorization Datasets, Maintained by: Evgeniy Gabrilovich gabr@cs.technion.ac.il
7. Davis, T.A.: The University of Florida Sparse Matrix Collection
8. Tewarson, R.P.: Sparse matrices. ELSEVIER
9. Liu, X., Yu, S., Moreau, Y., De Moor, B., Glänzel, W., Janssens, F.: Hybrid Clustering of Text Mining and Bibliometrics Applied to Journal Sets. In: Siam Proceeding on Data Mining (2009)
10. Arnold, G., Holzl, J., Koksal, A.S., Bodík, R., Sagiv, M.: Specifying and Verifying Sparse Matrix Codes. In: ICFP 2010 Proceedings of the 15th ACM SIGPLAN International Conference on Functional Programming (2010)

Unconstrained Optimization for Maximizing Ultimate Tensile Strength of Pulsed Current Micro Plasma Arc Welded Inconel 625 Sheets

Kondapalli Siva Prasad¹, Y.V. Srinivasa Murthy², Ch. Srinivasa Rao³,
D. Nageswara Rao⁴, and Gurralla Jagadish²

¹ Department of Mechanical Engineering, Anil Neerukonda Institute of Technology & Sciences, Visakhapatnam, 531 162, India

kspanits@gmail.com

² Department of Computer Science & Engineering, Anil Neerukonda Institute of Technology & Sciences, Visakhapatnam, 531 162, India

{urvishnu, jagadish1215}@gmail.com

³ Department of Mechanical Engineering, Andhra University, Visakhapatnam, 530 002, India

⁴ Centurion University, India

Abstract. Nickel alloys had gathered wide acceptance in the fabrication of components which require high temperature resistance and corrosion resistance. The paper focuses on developing mathematical model to predict ultimate tensile strength of pulsed current micro plasma arc welded Inconel 625 nickel alloy. Four factors, five level, central composite rotatable design matrix is used to optimize the number of experiments. The mathematical model has been developed by response surface method and its adequacy is checked by ANOVA technique. By using the developed mathematical model, ultimate tensile strength of the weld joints can be predicted with 99% confidence level. Contour plots are drawn to study the interaction effect of welding parameters on ultimate tensile strength of Inconel 625 weld joints. The developed mathematical model has been optimized using Hooke and Jeeves Method to maximize the ultimate tensile strength.

1 Introduction

In welding processes, the input parameters have greater influence on the mechanical properties of the weld joints. By varying the input process parameters, the output could be changed with significant variation in their mechanical properties. Accordingly, welding is usually selected to get a welded joint with excellent mechanical properties. To determine these welding combinations that would lead to excellent mechanical properties, different methods and approaches have been used. Various optimization methods can be applied to define the desired output variables through developing mathematical models to specify the relationship between the input parameters and output variables. One of the most widely used methods to solve this problem is response surface methodology (RSM), in which the unknown mechanism

with an appropriate empirical model is approximated, being the function of representing a response surface method.

Pulsed current MPAW involves cycling the welding current at selected regular frequency. The maximum current is selected to give adequate penetration and bead contour, while the minimum is set at a level sufficient to maintain a stable arc [1,2]. This permits arc energy to be used effectively to fuse a spot of controlled dimensions in a short time producing the weld as a series of overlapping nuggets. There are four independent parameters that influence the process are peak current, back current, pulse and pulse width.

From the literature review [3-8] it is understood that in most of the works reported the effect of welding current, arc voltage, welding speed, wire feed rate, magnitude of ion gas flow, torch stand-off, plasma gas flow rate on weld quality characteristics like front melting width, back melting width, weld reinforcement, welding groove root penetration, welding groove width, front-side undercut are considered. However much effort was not made to develop mathematical model to predict the same especially when welding thin sheets in a flat position. Hence an attempt is made to correlate important pulsed current MPAW process parameters to ultimate tensile strength of the weld joints by developing mathematical model and optimizing using Hooke and Jeeves method for pulsed current MPAW welded Inconel625 sheets.

2 Experimental Procedure

Inconel625 sheets of 100 x 150 x 0.25mm are welded autogenously with square butt joint without edge preparation. High purity argon gas (99.99%) is used as a shielding gas and a trailing gas right after welding to prevent absorption of oxygen and nitrogen from the atmosphere. From the literature four important factors of pulsed current MPAW as presented in Table 1 are chosen. All other parameters are kept constant. A large number of trial experiments are carried out using 0.25mm thick Inconel625 sheets to find out the feasible working limits of pulsed current MPAW process parameters. Due to wide range of factors, it was decided to use four factors, five levels, rotatable central composite design matrix to perform the number of experiments for investigation. Table 2 indicates the 31 set of coded conditions used to form the design matrix. The first sixteen experimental conditions (rows) have been formed for main effects. The next eight experimental conditions are called as corner points and the last seven experimental conditions are known as center points. The method of designing such matrix is dealt elsewhere [9,10]. For the convenience of recording and processing the experimental data, the upper and lower levels of the factors are coded as +2 and -2, respectively and the coded values of any intermediate levels can be calculated by using the expression [11].

$$X_i = 2[2X - (X_{\max} + X_{\min})] / (X_{\max} - X_{\min}) \quad (1)$$

Where X_i is the required coded value of a parameter X . The X is any value of the parameter from X_{\min} to X_{\max} , where X_{\min} is the lower limit of the parameter and X_{\max} is the upper limit of the parameter.

Table 1. Important factors and their levels

SI No	Input Factor	Units	Levels				
			-2	-1	0	+1	+2
1	Peak Current	Amps	6	6.5	7	7.5	8
2	Back Current	Amps	3	3.5	4	4.5	5
3	Pulse	No's/sec	20	30	40	50	60
4	Pulse width	%	30	40	50	60	70

Table 2. Design matrix and experimental results

SI No	Peak Current (Amps)	Back current (Amps)	Pulse (No/sec)	Pulse width (%)	Ultimate tensile strength(UTS) (MPa)
1	-1	-1	-1	-1	833
2	1	-1	-1	-1	825
3	-1	1	-1	-1	838
4	1	1	-1	-1	826
1	-1	-1	1	-1	826
2	1	-1	1	-1	830
7	-1	1	1	-1	825
8	1	1	1	-1	826
9	-1	-1	-1	1	825
10	1	-1	-1	1	820
11	-1	1	-1	1	835
12	1	1	-1	1	828
13	-1	-1	1	1	818
14	1	-1	1	1	826
11	-1	1	1	1	824
12	1	1	1	1	830
17	-2	0	0	0	830
18	2	0	0	0	826
19	0	-2	0	0	821
20	0	2	0	0	828
21	0	0	-2	0	832
22	0	0	2	0	825
23	0	0	0	-2	831
24	0	0	0	2	825
21	0	0	0	0	830
22	0	0	0	0	830
27	0	0	0	0	840
28	0	0	0	0	830
29	0	0	0	0	838
30	0	0	0	0	830
31	0	0	0	0	834

Tensile tests are carried out in 100KN computer controlled Universal Testing Machine (ZENON, Model No: WDW-100). The specimen is loaded at a rate of 1.5KN/min as per ASTM specifications and the values of ultimate tensile strength of the weld joints was evaluated and the results are presented in Table 2.

3 Developing Mathematical Model

The ultimate tensile strength of the weld joint is a function of peak current (A), back current (B), pulse (C) and pulse width (D). It can be expressed as [12-14].

Ultimate tensile strength (T)

$$T = f(A, B, C, D) \tag{2}$$

Using MINITAB 14 statistical software package, the significant coefficients were determined and final model is developed using significant coefficients to estimate ultimate tensile strength values of weld joint.

The final mathematical model are given by Ultimate tensile strength (T)

$$T = 833.143 - 0.875X_1 + 1.792X_2 - 1.625X_3 - 1.458X_4 - 1.296X_1^2 - 2.171X_2^2 - 1.296X_4^2 + 3.187X_1X_3 \tag{3}$$

Where X_1, X_2, X_3 and X_4 are the coded values of peak current, back current, pulse and pulse width.

4 Effect of Process Variables on Output Response

Contour plots play a very important role in the study of the response surface. Fig's 1a to 1b represents the contour plots for ultimate tensile strength. From the contour plots, the interaction effect between the input process parameters and output response can be clearly analyzed.

From the contour plot in Fig.1a, for optimum ultimate tensile strength of pulsed current MPAW Inconel 625 nickel alloy, the tensile strength is more sensitive to change in peak current than in the back current. From the contour plot in Fig.1b, it can be seen that ultimate tensile strength is more sensitive to pulse than pulse width. From all the contour plots it is understood that peak current and pulse are the most important parameters which affect the ultimate tensile strength of the welded joints.

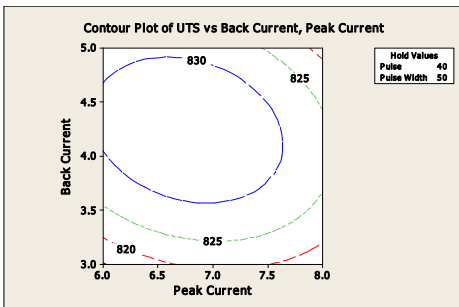


Fig. 1a.

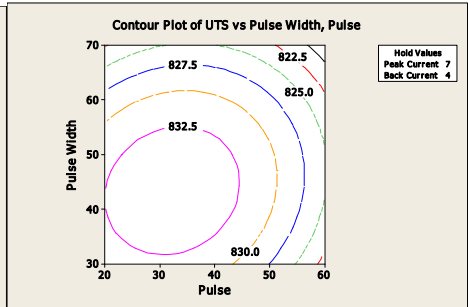


Fig. 1b.

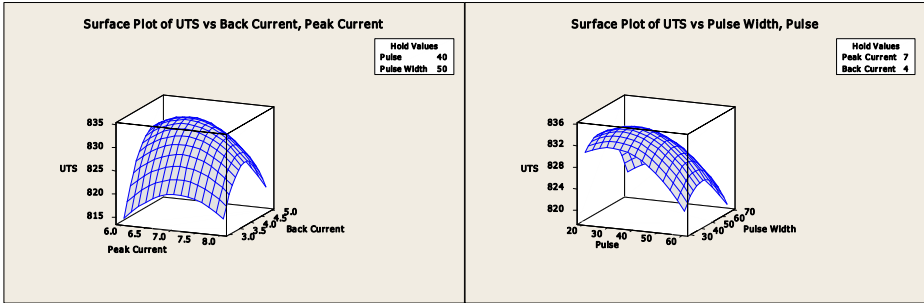


Fig. 2a.

Fig. 2b.

Response Surface plots clearly indicate the optimal response point. The optimum ultimate tensile strength of pulsed current MPAW welded Inconel625 nickel alloy was exhibited by the apex of the response surface, as shown in Fig.2a & 2b

Fig.2a shows the three dimensional response surface lot for ultimate tensile strength obtained from the regression model, assuming a peak current of 7 Amps and a back current of about 4.1Amps. The optimum ultimate tensile strength is exhibited by the apex of the response surface. From the response graph, it is identified that at the peak current of 7 Amps, the ultimate tensile strength of pulsed current MPAW joints is higher. The formation of fine equiaxed grains in fusion zone increases the ultimate tensile strength of the welded joints. When peak current is increased from 7 Amps, the ultimate tensile strength decreases. This is the result of the increased heat input associated with the use of higher peak current.

Fig.2b shows the three dimensional response surface plot for the response ultimate tensile strength obtained from the regression model, assuming pulse of 40 pulses/sec and pulse width of 10%. From the response graph, it is observed that when pulse is 40pulse/sec, the tensile strength of the pulsed current MPAW welded joint is higher. The fine grains observed in the fusion zone due to optimum heat input may be responsible for the better ultimate tensile strength of these welded joints.

5 Optimization of Pulsed Current MPAW Process Parameters

Hooke & Jeeves method [16] is used to optimization tool to search the optimum values of the process variables. In this paper the algorithm is developed to optimize the pulsed current MPAW process variables. The objective is to maximize ultimate tensile strength.

The Hooke & Jeeves method incorporates the past history of a sequence of iterations into the generation of a new search direction. It combines exploratory moves with pattern moves. The exploratory moves examine the local behavior of the function & seek to locate the direction of any stepping valleys that might be present. The pattern moves utilize the information generated in the exploration to step rapidly along the valleys.

5.1 Exploratory Move

Given a specified step size which may be different for each co-ordinate direction and change during search is done. The exploration proceeds from an initial point by the specified step size in each coordinate direction. If the function value does not increased the step is considered successful. Otherwise the step is retracted and replaced by a step in the opposite direction which in turn is retained in depending upon whether it success or fails. When all N coordinates have been investigated, the exploration move is completed. The resulting point is termed a base point.

5.2 Pattern Move

A pattern move consists of a single step from the present base point along the line from the previous to the current base point.

A new pattern point is calculated as:

$$x_p^{(k+1)} = x^{(k)} + (x^{(k)} - x^{(k-1)})$$

where, $x_p^{(k+1)}$ is temporary base point for a new exploratory move.

If the result of this exploration move is a better point then the previous base point ($x^{(k)}$) then this is accepted as the new base point $x^{(k+1)}$. If the exploratory move does not produce improvement, the pattern move is discarded and the search returns to $x^{(k)}$, where an exploratory search is undertaken to find a new pattern.

Steps:

Step 1: Starting pint $x^{(0)}$

The increments Δ_i for $i=1,2,3, \dots, N$

Step reduction factor $\alpha > 1$

A termination parameter $\epsilon > 0$

Step 2: Perform exploratory search

Step 3: Was exploratory search successful (i.e. was a lower point found)

If Yes go to step (1)

Else continue

Step 4: check for the termination $\|\Delta\| < \epsilon$ current pint approximation x_0

$$\Delta_i = \Delta_i / \alpha \text{ for } i = 1,2,3, \dots, N$$

Go to step 2

Step 1: Perform pattern move

$$x_p^{(k+1)} = x^{(k)} + (x^{(k)} - x^{(k-1)})$$

Step 2: Perform exploratory research using $x_p^{(k+1)}$ as the base point; let the result be $x^{(k+1)}$.

Step 7: Is $f(x^{(k+1)}) < f(x^{(k)})$?

If Yes Set $x^{(k-1)} = x^{(k)}$

$x^{(k)} = x^{(k+1)}$ go to step (1).

Else go to step (4)

Table 3. Optimized pulsed current MPAW Parameters

Parameter	Hooke & Jeeves Method	Experimental
Peak current(Amps)	7.2177	7
Back current(Amps)	4.2177	4
Pulse (no/sec)	44.3545	40
Pulse width(%)	54.3545	50
Maximum ultimate tensile strength(Mpa)	844.3545	840

From Table 3 it is understood that the values predicted by Hooke and Jeeves method and experimental values are very close to each other.

6 Conclusions

Empirical relation is developed to predict ultimate tensile strength of pulsed current micro plasma arc welded Inconel 625 nickel alloy using response surface method. The developed model can be effectively used to predict ultimate tensile strength of pulsed current micro plasma arc welded joints. From the contour plots, it is understood that peak current and pulse are the more sensitive to changes in ultimate tensile strength of the welded joint than back current and pulse width. From the experiments conducted it is observed that maximum ultimate tensile strength obtained is 840 Mpa for the input parameter combination of peak current of 7Amps, back current of 4 Amps, pulse of 40 pulses /sec and pulse width of 50%. From Hooke and Jeeves method the maximum value obtained is 844.3545 MPa for the input parameter combination of peak current of 7.2177Amps, back current of 4.2177 Amps, pulse of 44.3545pulses /sec and pulse width of 54.3545%.The values obtained experimentally and predicted by Hooke and Jeeves method are very close to each other.

Acknowledgments. The authors would like to thank Shri. R.Gopla Krishnan, Director, M/s Metallic Bellows (I) Pvt Ltd, Chennai, India and Dr. S. C Satapathy, CSE Department, ANITS for their support to carry out experimentation and programming work.

References

1. Balasubramanian, B., Jayabalan, V., Balasubramanian, V.: Optimizing the Pulsed Current Gas Tungsten Arc Welding Parameters. *J. Mater Sci. Technol.* 22, 821–821 (2002)
2. Madusudhana Reddy, G., Gokhale, A.A., Prasad Rao, K.: Weld microstructure refinement in a 1441 grade aluminium-lithium alloy. *Journal of Material Science* 32, 4117–4122 (1997)
3. Zhang, D.K., Niu, J.T.: Application of Artificial Neural Network modeling to Plasma Arc Welding of Aluminum alloys. *Journal of Advanced Metallurgical Sciences* 13, 194–200 (2000)

4. Chi, S.-C., Hsu, L.-C.: A fuzzy Radial Basis Function Neural Network for Predicting Multiple Quality characteristics of Plasma Arc Welding. *IEEE*, pp. 2807–2812 (2001) 0-7803-7078-3/01
5. Hsiao, Y.F., Tarng, Y.S., Wang, J.: Huang Optimization of Plasma Arc Welding Parameters by Using the Taguchi Method with the Grey Relational Analysis. *Journal of Materials and Manufacturing Processes* 23, 11–18 (2008)
6. Siva, K., Muragan, N., Logesh, R.: Optimization of weld bead geometry in Plasma transferred arc hardfacing austenitic stainless steel plates using genetic algorithm. *Int. J. Adv. Manuf. Technol.* 41, 24–30 (2008)
7. Lakshinarayana, A.K., Balasubramanian, V., Varahamoorthy, R., Babu, S.: Predicted the Dilution of Plasma Transferred Arc Hardfacing of Stellite on Carbon Steel using Response Surface Methodology. *Metals and Materials International* 14, 779–789 (2008)
8. Balasubramanian, V., Lakshminarayanan, A.K., Varahamoorthy, R., Babu, S.: Application of Response Surface Methodology to Prediction of Dilution in Plasma Transferred Arc Hardfacing of Stainless Steel on Carbon Steel. *Science Direct* 12, 44–53 (2009)
9. Montgomery, D.C.: *Design and analysis of experiments*, 3rd edn., pp. 291–291. John Wiley & Sons, New York (1991)
10. BoxG, E.P., Hunter, W.H., Hunter, J.S.: *Statistics for experiments*, pp. 112–111. John Wiley & Sons, New York (1978)
11. Ravindra, J., Parmar, R.S.: Mathematical model to predict weld bead geometry for flux cored arc welding. *Journal of Metal Construction* 19, 12–41 (1987)
12. Cochran, W.G., Cox, G.M.: *Experimental Designs*. John Wiley & Sons Inc., London (1957)
13. Barker, T.B.: *Quality by experimental design*. ASQC Quality Press, Marcel Dekker (1981)
14. Gardiner, W.P., Gettinby, G.: *Experimental design techniques in statistical practice*. Horwood press, Chichester (1998)
15. Kalyanmoy, D.: *Optimization for engineering design*. Prentice Hall (1988)

Simple and Effective Techniques for Skew Correction, Slant Correction and Core-Region Detection for Cursive Word Recognition

Kota Virajitha¹, B. Navya¹, L.N. Phaneendra Kumar Boggavarapu¹,
Radhe Syam Vaddi¹, and Hima Deepthi Vankayalapati²

¹ Department of Information Technology, V R Siddhartha Engineering College,
Vijayawada, India

² Department of Computer Science & Engineering, V R Siddhartha Engineering
College, Vijayawada, India

Abstract. For the past decades, the advancement in the field of Image Processing has been paving a profound way in digital treatment of Human written data. Handwriting Recognition, a subset, is now a major research area to study as it is providing a mean for automatic processing of large volumes of data in reading and office automation. Intelligent word recognition systems which are used in processing important documents like bank cheques, old scripts are the need of the hour. Through this paper we present a new approach for Cursive word and Signature recognition. We propose Core-region detection technique which enables us to identify the crucial features of the hand written signatures by the extracting ‘Ascenders and Descenders’. Skew and Slant corrections, if needed, are performed as preprocessing steps. A significant reduction in computation complexity has been observed than the previous attempts of researchers in detection of core-region.

1 Introduction

Cursive word recognition is growing fast and getting more vital for the past six years. Our paper concentrates on Feature extraction (Extraction of ascenders and Descenders). This paper is a collective approach to demonstrate prominent Preprocessing techniques (Skew and Slant correction) and Feature Extraction. Each of the areas in a Word recognition system is dealt with great detail in many previous attempts but this paper is one of the few attempts which collectively deal with both Preprocessing and Feature Extraction.

Many methods have been developed in an attempt to satisfy the need for Word recognition systems that exists in various applications like automatic reading of postal addresses and bank checks, processing- documents such as forms, etc. The typical modules of a Word recognition system are preprocessing, then a possible segmentation or fragmentation phase, feature extraction, the core of recognition, and post-processing. *Preprocessing* usually includes normalization, noise reduction, reference line finding, and either contour or skeleton tracing if necessary. The preprocessing starting point depends on the environment in which the system is

running. It may include external word segmentation (extraction) from a multi-word neighborhood and other various document processing techniques. Given a stand-alone word, a few normalization operations are performed, among which are:

1. *Skew correction*: a rotation transformation that brings the word orientation parallel to the horizontal;
2. *Slant correction*: a shear transformation that attempts to make all the vertical strokes erect;
3. *Smoothing*: includes all different kinds of noise reduction.
4. *Scaling*: invariance to size (used in rare cases only).

Next, there is the *Segmentation* phase and its substitutes. In a segmentation process, in contrast with simple fragmentation or splitting into pieces, there is an attempt to split the word image into segments that relate to characters. *Feature extraction* process takes place next. When high resolution features are used, the extraction process is more sensitive to noise. The objective of this stage is to retrieve observations out of the word image. There are several classes of features. Segmentation-free methods use either raw features, which are pixel-wise like strokes, or global symbolic features such as ascenders, descenders, loops, etc. After the *Recognition module* has finished running, some methods use *Post-processing* techniques to improve the recognition results.

As mentioned earlier through our paper we present our approach towards Feature Extraction .i.e. (Extraction of Ascenders and Descenders).The Skew and Slant corrected word images are processed to extract distinct features of the word. Ascenders and Descenders are very prominent features used by many of the word recognition techniques. An Ascender is the part of the word that extends above the top reference line of the word (Top Line) and Descender is that part of the word which extends below the bottom reference line (Base Line). The region between the Top line and the Base line is called 'Core-region'.

2 Related Work

The cursive word recognition strategies reported in literature employ different methodologies for Preprocessing and Extraction of ascenders/descenders[1-2]. Bozinovic and Srihari [3] are the first who were succeeded in detecting the reference lines based on horizontal density histogram (popular as BSM).In order to find the actual core region lines their technique needed to evaluate lots of heuristic rules[4]. Some alternative techniques were proposed in [5].Here, rather than using the density histogram, mere analysis of density distribution declined the influence of local strokes. However, the reason that the core-region was erroneously detected due to the fact that upper baseline and lower baseline were incorrectly set. This occurred due to presence of erratic characters and multiple characters that contained long horizontal strokes such as the letter "t", "f", "h" etc. in the string [9].A weighted average based combination of individual component slants to normalize a numeral string is proposed in [6]. A new de-skewing algorithm based on the entropy computation of the image is presented in [7]. Radon Transform based algorithms for skew and slant corrections are presented in [11]. The Radon transform of the image and its gradient is used to estimate the skew angle of the word image, long strokes and thier average angle for

the estimation of slant angle. Though Radon transform produce results, it suffers from high computation complexity. A Cellular Neural Network (CNN) based algorithm to detect lower and upper base lines of a word is presented in [13]. The ascenders and descenders are further located by detecting patches of the word images in regions above and below the upper and lower base lines in [12] and by using the outer contour of the line and height thresholds is presented in [14].

On the other hand, literature is abundant with several approaches to slant angle detection, which can be roughly divided into three groups. The first two groups deal with uniform slant correction while the third group deals with local slant correction. Bozinovic and Srihari, calculate slant angle by detecting near vertical strokes and taking the average angle of those as the shear angle. Later, several others such as Kim and Govindaraju [15]; Vinciarelli and Luettin [7]; Shridar and Kimura [16] have followed their lead, using different ways to select the strokes. For any given image, the text is sheared to discrete number of angles around the vertical orientation. For each of these images, the vertical projection profile is calculated; among them whichever has maximum variation is taken as the profile. In comparison to the other approaches this method is quite fast, but on the downside, it relies heavily on heuristics, hence is not very robust. In addition, the accuracy of such approaches requires accurate detection of the edges of the characters composing the word. The second approach evaluates a measure function of the image on a range of shear angles and selects the angle with highest value of the measure. Finally, the hybrid approach of Wigner-Ville distribution and projection profile technique was integrated in a complete image processing system [17]. Likewise, few more approaches are detailed in [8]. However, such methods are computationally heavy since multiple shear transformed word images corresponding to different angles in an interval have to be calculated in addition to the slant angle based on structure features of all characters.

The third approach distinguishes itself from the first two by correcting the slant non-uniformly. The techniques above shear a word (or bigger units) uniformly, i.e. by a single angle, hence can never fully cope with variant-slanted words. The approaches presented in [10, 12] are used to handle local slant (non-uniform) by employing dynamic programming techniques. To apply different shear angles at different points within a word, one has to split the word up into intervals and shear each of those individually. To determine what intervals to take, and by what angle to shear over each interval, a criterion is optimized that evaluates the sequences of intervals and angles simultaneously. Such a method has a lot of potential, since it can cope with variant-slanted words. Indeed, the results are encouraging, but on the other hand, there are more robustness issues, as the algorithm has greater freedom to make errors within a word. Additionally, theoretical background and mathematical techniques are somewhat more demanding. Finally, an independent correction of each component is also not practicable, since this may produce distortions when broken characters are present in the string.

Most of these methodologies use complex models of word images and transforms for achieving tilt corrections of the word and for extraction of ascenders and descenders. Keeping in view the merits and demerits of the previous attempts discussed above, we designed our approach to provide profitable results. In the next section, we will evaluate the approach in order to detect the core-region. In further

sections we described Skew- detection/correction; Slant detection/correction; Core-region detection in detail. Finally we will coin the merits of work and future work of our approach in the last section.

3 Proposed Method

3.1 Identification of Core Region

Identification of Core Region for a given input signature involves identifying the featured crucial area within the given input signature or script. This entire process is organized into three categories depending on the functionality needed.

1. *Skew Estimation and Correction:* At this step, we obtain the skew angle and correct the signature by this skew angle. The output of this step is to get the signature parallel to horizontal axis.

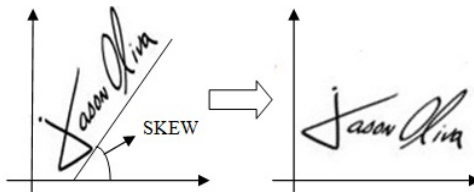


Fig. 1(a). Image with Skew Fig. 1(b). Skew Corrected image

2. *Slant Estimation and Correction:* In this step, we obtain the slant angle and correct the signature by this slant angle. De-skewed image (Image free from skew) is given an input and signature parallel to vertical axis is obtained as output.

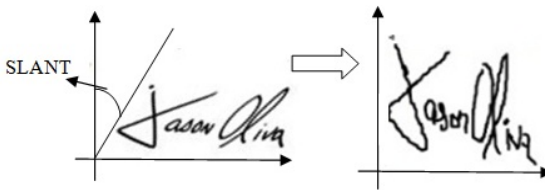


Fig. 2(a). Image with Slant Fig. 2 (b). Slant Corrected image

3. *Extract Ascenders and Descenders:* Here, we obtain the Top line and the Base line. The characters above the top line are called 'Ascenders' and those below the base line are called the 'Descenders'.

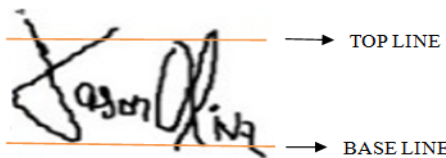


Fig. 3. Extraction of Ascenders and Descenders

The resultant Flow chart of the proposed method is an integration of above three procedures. The flow is as shown in the figure 4.

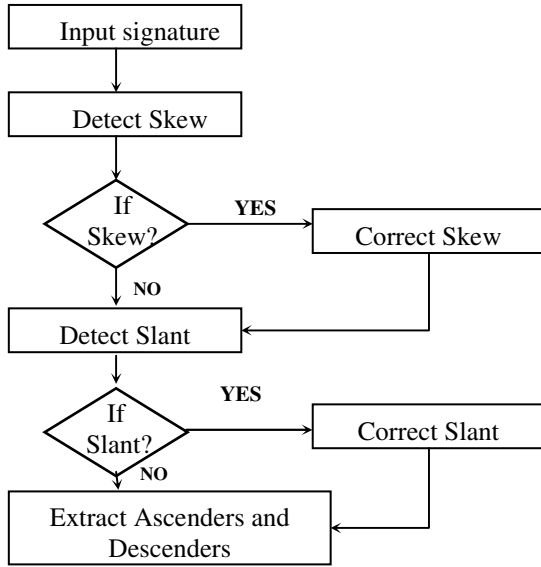


Fig. 4. Flow chart depicting the Proposed Method

4 Implementation

4.1 Skew Detection and Correction

- 1] First load the signature. Change the image to black and white format and thin the image. This is used to remove the noise, and to get the accurate values.
- 2] Extract the number of foreground pixels.
- 3] Move the signature to origin by using the co-ordinates of center of mass of the signature image.
- 4] Calculate the minimum Eigen value of the matrix formed by using the new co-ordinates of the signature image.
- 5] Calculate Skew angle using the Eigen vector.

$$\phi = \left[\frac{M(1)}{M(2)} \right] \quad (1)$$

Where ϕ is the Skew angle;

M is the image matrix;

M(1) is the first element in the first row of the image matrix;

M(2) is the first element in the second row of the image matrix.

- 6] After Skew angle is found, Skew correction is performed by applying rotation transformation to every pixel (foreground) in the image.



Fig. 5. Signature given as an input

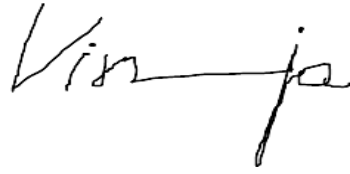


Fig. 6. Deskewed image obtained as output

4.2 Slant Detection and Correction

The Deskewed image obtained above is given as input to this algorithm.

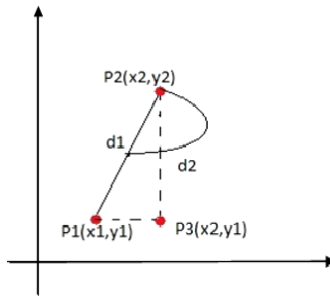


Fig. 7. Illustration of Slant angle estimation

- 1] Obtain the left most pixel value as $P_1=(x_1, y_1)$.
- 2] Next get the first maxima from leftmost as $P_2=(x_2, y_2)$.
- 3] Set the value of $P_3=(x_2, y_1)$.
- 4] Calculate the distance between P_1 and P_2 . Let it be ' d_1 '.
- 5] Distance between P_2 and P_3 as ' d_2 '.
- 6] Find the slope ' m ' between lines formed by joining P_1, P_2 and P_2, P_3 . If $m < 0$ then $k = -1$ else $k = 1$
- 7] Now Slant angle is calculated by using: $\theta = k \sin^{-1}(d_2 / d_1)$ (2)
- 8] For every foreground pixel in the signature slant correction is performed by applying the transformation:

$$x_1 = x - y \tan(90 - \theta) \quad \text{and} \quad y_1 = y \quad (3)$$



Fig. 8. Plotting leftmost pixel



Fig. 9. Plotting Leftmost pixel from top



Fig. 10. Deslant image obtained as output

4.3 Extraction of Ascenders and Descenders

The Deskewed and Deslant image from the above processes is passed as an input to this process. The procedure is as follows:

- 1] Initially obtain the number of foreground pixels in each row (a_i).
- 2] Now, obtain the middle line as the index value where there are maximum no of pixels (i.e.) $\text{mid} = \text{index of max}(a_i)$.
- 3] Sort all the a_i values and name this sorted list as a_j
- 4] Calculate third quantile Q_3 .

$$Q_3 = a_j, \quad j = \lceil h \times 3/4 \rceil$$

The value of index (j) is $3/4$ of the height of the image. We obtain the value in the a_j list with the above index value. We set this value as Q_3 .

- 5] Trace number of pixels from first row to middle line. If the obtained pixel value is greater than or equal to Q_3 the set this index value of the row as the Top line.
- 6] Trace number of pixels from maximum height to middle line. If the obtained pixel value is greater than or equal to Q_3 the set this index value of the row as the Baseline.

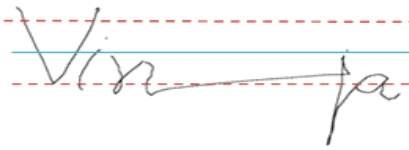


Fig. 11. Extraction of Ascenders, Descenders and Core region Detection

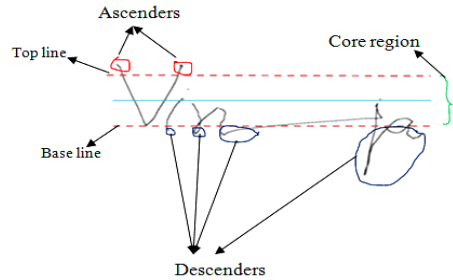


Fig. 12. Depiction of Ascenders, Descenders and Core-region

Thus we obtained the final output (i.e.) Feature extraction– Extraction of Ascenders and Descenders.

5 Conclusion

Through this paper we presented simple techniques for tilt correction and extraction of ascender/descender features from the word image. The algorithms for tilt-correction

and extraction of ascenders/ descenders have been implemented in MATLAB and are tested on a large sample of signature images. The performance of each of the tilt correction and ascender/descender extraction algorithms were evaluated by visual inspection of the results. Enhanced technique for core-region detection exhibits better results generally and particularly for erratic words. Simplicity of the proposed slant correction approach reduced computational complexity significantly. Hence, it avoided heavy experimental efforts required to find the optimal configuration of a parameter set. Moreover, long exploration of the parameter space is avoided.

Few challenges remain during the treatment of words that include characters with different slants. Consequently, we are concentrating to update our approach to deal with non-uniform slanted words.

References

1. Chin, W., Harvey, M., Jennings, A.: Skew Detection in Hand written Scripts. In: IEEE TENCON, Speech and Image Technologies for Computing and Telecommunications (1997)
2. Kava Ieratou, E., Fakotakis, N., Konakis, G.: New Algorithms for Skew Correction and Slant Removal on Word level. Proceedings of 6th IEEE International Conference on Electronics, Circuits and Systems 2, 1159–1162 (1999)
3. Bozinovic, R., Srihari, S.: Off-line Cursive Script Word Recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence 11(1), 68–83 (1989)
4. Morita, M., Facon, J., Bortolozzi, F., Garnes, S., Sabourin, R.: Mathematical morphology and weighted least squares to correct handwriting baseline skew. In: Proceedings of the International Conference on Document Analysis and Recognition, vol. 1, pp. 430–433 (1999)
5. Cote, M., Lecolinet, E., Cheriet, M., Suen, C.: Automatic reading of cursive scripts using a reading model and perceptual concepts. International Journal on Document Analysis and Recognition 1(1), 3–17 (1998)
6. Britto Jr., A.D.S., Robert, S., Edouard, L., Flavio, B.: Improvement in hand written Numeral String Recognition By Slant Normalization and Contextual Information. In: Proceedings of SIWFHR, Amsterdam, September 11-13, pp. 323–332 (2000)
7. Alessandro, V., Luetin, J.: A New Normalization Technique for Cursive Handwritten Words. Pattern Recognition Letters 22, 1043–1050 (2001)
8. El-Yacoubi, A., Gilloux, M., Sabourin, R., Suen, C.Y.: An HMM-based Approach for on-line unconstrained handwritten word modeling and recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 21(8), 752–760 (1999)
9. Blumenstein, M., Cheng, C.K., Liu, X.Y.: New preprocessing techniques for handwritten word recognition
10. Taira, E., Uchida, S., Sakoe, H.: Non-uniform slant correction for handwritten word recognition. IEICE Transactions on Information & Systems E87-D(5), 1247–1253 (2004)
11. Dong, J.-X., Krzyżak, D.P.A., Suen, C.Y.: Cursive word skew/slant corrections based on Radon transform. In: Proceedings of International Conference on Document Analysis and Recognition (ICDAR), pp. 478–483 (2005)
12. Seiichi, U., Eiji, T., Hiroaki, S.: Non Uniform Slant Correction using Dynamic Programming. In: Proceedings of International Conference on Document analysis and Recognition, ICDAR (2001)

13. Kristof, K., Gabor, P., Tamas, R.: Intimate Integration of Shape codes and Linguistic Framework in Handwriting Recognition via Wave Computers. In: Proceedings of 16th European Conference on Circuits Theory and Design, Poland, pp. 409–412 (2003)
14. Didier, G., Suen, C.Y.: Cursive Script Recognition Applied to Processing of Bank Cheques. In: Proceedings of International Conference on Document Analysis and Recognition (ICDAR), pp. 11–14 (1995)
15. Kim, G., Govindaraju, V., Ecient: Chain-code-based image manipulation for handwritten word recognition. In: Proceedings of SPIE-The International Society for Optical Engineering, Bellingham, WA, USA, vol. 2, pp. 262–272 (1996)
16. Shridar, M., Kimura, F.: Handwritten address interpretation using word recognition with and without lexicon. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Piscataway, NJ, USA, vol. 3, pp. 2341–2346 (1995)
17. Kavallieratou, E., Dromazou, N., Fakotakis, N., Kokkinakis, G.: An Integrated System for Handwritten Document Image Processing. *International Journal on Pattern Recognition and Artificial Intelligence* 17(4), 617–636 (2003)

A New CDMA Based 2.5G Network Base Station Power Control Algorithm for Improving User Capacity

Ramarakula Madhu and G. Sasi Bhushana Rao

Department of Electronics and Communication Engineering,
College of Engineering, Andhra University, Visakhapatnam-530003, A.P, India
madhu_ramarkula@rediffmail.com, sasi_gps@yahoo.co.in

Abstract. In TDMA and FDMA each user is allocated a unique time and frequency slot. The capacity of these systems has a predefined value; therefore power control has no effect on capacity. In CDMA every mobile user is allowed to transmit on the full available spectrum simultaneously. In this the capacity is defined by the number of users that can be accommodated without losing acceptable QoS. The SIR of all users should not fall below a threshold, so power control in CDMA has a great impact on capacity. In this paper, a new BS Power Control algorithm is proposed for 2.5G network to achieve an average SIR for all users which is above a certain threshold. This algorithm is based on three-valued power update coefficients (≥ 5 , between 2 and 5; and ≤ 2) and implemented for three static users of a cell at every epoch with 16 iterations for improving user capacity.

Keywords: TDMA (Time Division Multiple Access), FDMA (Frequency Division Multiple Access), CDMA (Code Division Multiple Access), QoS (Quality of Service), SIR (Signal-to-Interference Ratio), BS (Base Station).

1 Introduction

Code Division Multiple Access (CDMA) is a type of a spread spectrum modulation technique [1], used in cellular system alternative to both Frequency Division Multiple Access (FDMA) and Time Division Multiple Access (TDMA). In FDMA strategies, the focus is on the frequency dimension. Here, the total bandwidth is divided into N narrowband frequency slots. So each user is allowed to communicate simultaneously by assigning the narrowband frequency slots. Since the total bandwidth subdivided into N frequency slots, only N users may be supported simultaneously. CDMA scheme allows multiple users to share the same frequency at all the time. Each user is identified by a code that is orthogonal to all other users since co-users are isolated by codes rather than frequencies.

In order to meet the increasing demand of mobile subscribers for various services, in 2.5G, such as web browsing and e-mail, it is crucial to have higher capacity and more severe Quality of Service (QoS) requirement. The standard 2.5G technology

which uses CDMA is the General Packet Radio Service (GPRS) and Enhanced Data rates for GSM Evolution (EDGE). Unlike FDMA and TDMA, CDMA has a soft capacity. This means that there is no hard limit to how many users we can allow on the system. CDMA has an advantage of having more system capacity than the other multiple access schemes. The capacity in CDMA is defined by the number of users that are maintaining a Signal-to-Interference Ratio (SIR) above a predefined threshold. In these systems the SIR has a major effect on the capacity. Therefore transmission power control is necessary in CDMA to increase the capacity. When power control is not implemented, all mobiles transmit their signal with the same power without taking the consideration of the distance from the base station, in this case mobile users close to the base station will cause a high level of interference to the mobile users that are far away from the base station, this problem is known as the near-far effect. The improper power control in CDMA systems causes a reduction of capacity by 50% or more [2].

Power control manages the problems of near-far mentioned above by constantly controlling the received power of the mobiles and continuously adjusting its transmitting power in order to achieve the threshold SIR.

Most of the practical power control algorithms present today require high number of iterations in order to reach required SIR among the mobile users. In this paper a new Base Station Power Control Algorithm is proposed based on the three-valued power update coefficients (≥ 5 , between 2 and 5, and ≤ 2), to maintain an optimum SIR of all mobile users of 2.5G. This base station power control algorithm is implemented for three static mobile users of a cell at the CDMA base station with 16 iterations.

2 Power Control

Power control is simply the technique of controlling the mobile user power so as to affect the base station received power and hence the overall SIR. With appropriate power control, the CDMA offers high capacity in comparison to FDMA and TDMA. Since CDMA systems do not have time or frequency restriction among users. Each user changes its access to the resources by adapting its transmitting power. Therefore power control is a significant design problem in CDMA.

The concept of power control method is to compare the received power of each signal with a threshold. When the received power of user is below the threshold the receiver requests that the transmitter increase its output power and vice versa. There are two major classes of power control algorithms [3].

- Centralized Power Control
- Distributed Power Control

In centralized power control, a network center can simultaneously compute the optimal power levels for all users in the network. However it requires measurement of all the link gains and the communication overhead between a network center and base stations. Since all calculations done at the network center, the complexity and delay in these systems is too high for practical systems.

In distributed power control, the power calculations are done within the cells by determining the transmitter power of users. The delay and complexity are much lower and more scalable as compared with the centralized power control. So distributed power control is more suitable for practical applications. The algorithm proposed in this paper is a distributed power control algorithm.

3 Power Control Techniques

There are three techniques used for power control. They are; Slow Power Control, Open Loop Power Control and Closed Loop Power Control.

The Slow power control is used for the downlink. With this method the BS periodically decreases its transmitter power until it receives a request from a mobile station. Every mobile station measures the errors in each signal. When the error rate reaches a threshold, the mobile station requests extra power from the BS, thus maintains the power optimization [7].

In Open loop power control at the mobile user, the mobile user senses the received signal strength of the pilot signal and can adjust its transmit power based on that. If the signal level of the pilot signal is very strong, it can be assumed that the mobile user is very close to the base station. Therefore, the mobile user transmit power level should be reduced. At base station, the base station decreases its transmit power level gradually and waits to receive the error rate message from the mobile user. If the error rate is beyond a specified level, the base station increases its transmit power level on the corresponding channel.

The closed loop power control involves both uplink and down link [4]. This power control compensates for the variations in power over time and consists of inner and outer loops. In the inner loop, the base station measures the SIR of each user and compares it with a threshold. Mobile stations change their transmission power according to control commands from the base station. The outer loop is executed after 16 iterations of the inner loop. In this loop the desired SIR is adjusted in order to achieve the threshold. The characteristics of a better power control algorithm are to have a low overhead, a rapid handoff and no additional hardware requirement.

4 Power Control Algorithm in CDMA

Because of non-uniform distribution of users in CDMA, different QOS is delivered to users in different regions. For maintaining the same QOS among the users in the network, the transmitter power control should be applied. In addition to power control the user distribution in the cell is approximately equal.

Previous power control approaches in CDMA system assumed that only one antenna and matched filter receivers are being used at the base stations and each user employs an SIR based power update where the user's power is multiplied by the ratio of its target SIR to its current SIR. i.e., for user 'i', the update is

$$P_i(m + 1) = \frac{\xi_i^*}{\xi_i(m)} P_i(m) \tag{1}$$

Where $P_i(m)$ and $\xi_i(m)$ are the power and SIR of user 'i' at iteration (m) and ξ_i^* is the SIR target of user 'i'.

Sato and Takeo [5] introduced an algorithm based on the measurement of all user SIRs and the mean SIR. By adjusting the pilot signal power and the minimum acceptable power of the received signal, the difference between the SIRs of all users in the network was decreased.

In the Hanly algorithm [6] the distribution of the mobile stations among cells was determined in order to minimize the required transmit power of the mobile stations.

These algorithms have the following disadvantages:

- High overhead required maintaining the minimum transmission power.
- More number of periodic calculations is required for each mobile station.
- The algorithm leads to equal SIRs in all the cells but there is no guarantee that the desired SIR is achieved.

A new base station power control algorithm is proposed in this paper to obtain the optimum power control in a CDMA system of 2.5G users. This is a distributed algorithm in which the required calculations are done in parallel at all BSs in the network. The algorithm proposed here is superior because no additional hardware is required in the system to obtain the power control. In this algorithm a three-valued power update coefficient is used and the power is estimated at the base station with 16 iterations.

In this paper the algorithm is implemented for 3 static users of a cell; the base station measures the SIR of mobile users and compares it to the average SIR. In this case mobile users close to the base station will have a high SIR as compared to the mobile users that are far away from the base station. The power control is obtained among the users such that, if the SIR of the mobile user is less than the average SIR, the base station commands the mobile user to increase its transmit power and if the SIR of the mobile user is greater than the average SIR, the base station commands the mobile user to decrease its transmit power.

The steps involved in the proposed algorithm are:

Step1: Calculation of SIR of the three mobile users in the cell. The SIR is calculated from the formula,

$$SIR (dB) = P_r (dBW) - P_n (dBW) \tag{2}$$

Step2: Calculation of the average SIR of the mobile users at the base station.

Step3: Comparison of SIR of the mobile user with the average SIR. If the SIR of mobile user is below or above the average SIR, the transmission power of the mobile user is updated using the formula

$$P_k(m+1) = P_k(m) + x_k(m)[SIR_{avg}(m) - SIR_k(m)] \tag{3}$$

where, m is the number of iterations, $P_k(m)$ is the power of the k^{th} mobile user in the j^{th} cell at the m^{th} iteration, $P_k(m+1)$ is the power of the k^{th} mobile user in the $(m+1)^{th}$ iteration, $SIR_{avg}(m)$ is the average SIR of the mobile users in the j^{th} cell. $SIR_k(m)$

is the SIR of k^{th} mobile user and $x_k(m)$ is the three-valued power update coefficient of the k^{th} user at m^{th} iteration.

The power update coefficient $x_k(m)$ is given by

$$x_k(m) = \begin{cases} x_1, & |SIR_{avg}(m) - SIR_k(m)| \geq 5 \\ x_2, & 2 < |SIR_{avg}(m) - SIR_k(m)| < 5 \\ x_3, & |SIR_{avg}(m) - SIR_k(m)| \leq 2 \end{cases} \quad (4)$$

The values of x_1 , x_2 and x_3 are used here are 0.5, 0.25 and 0.125 respectively.

Step 4: Calculation of the updated transmission power of the mobile user at the end of the 16th iteration.

5 Results and Discussion

The three static mobile users; userA, userB and userC, with a distance of d_1 , d_2 and d_3 from the base station, are assumed. The relation between the distance of the mobile users is $d_3 < d_1 < d_2$. The noise power of all the three users is assumed as -71.76 dBW. The received power of UserA, UserB and UserC at the base station is -60 dBW, -64 dBW and -57 dBW respectively.

From equation (2), the SIR of userA is 15, the SIR of userB is 6 and the SIR of userC is 30. The average SIR (SIR_{avg}) of the users in the cell is 17. The transmit power of the mobile users is calculated for 5th, 10th and 16th iteration.

User A: The SIR of the *userA* is '15', which is less than the average SIR. Then $|SIR_{avg}(m) - SIR_k(m)| = 2$; therefore from equation (4) the update coefficient is $x_3 = 0.125$.

The transmit power of userA at the end of the 5th iteration is calculated as,

$$P_k(m+1) = -60 + 5 \times 0.125 [2] = -59.03 \text{ dBW}$$

The transmit power of userA at the end of the 10th iteration is calculated as,

$$P_k(m+1) = -60 + 10 \times 0.125 [2] = -56.02 \text{ dBW}$$

The transmit power of userA at the end of the 16th iteration is calculated as,

$$P_k(m+1) = -60 + 16 \times 0.125 [2] = -53.98 \text{ dBW}$$

User B: The SIR of the *userB* is '6', which is less than the average SIR. Then $|SIR_{avg}(m) - SIR_k(m)| = 11$; therefore from equation (4) the update coefficient is $x_1 = 0.5$.

The transmit power of userB at the end of the 5th iteration is calculated as,

$$P_k(m+1) = -64 + 5 \times 0.5 [11] = -49.6 \text{ dBW}$$

The transmit power of userB at the end of the 10th iteration is calculated as,

$$P_k(m + 1) = -64 + 10 \times 0.5[11] = -46.6 \text{ dBW}$$

The transmit power of userB at the end of the 16th iteration is calculated as,

$$P_k(m + 1) = -64 + 16 \times 0.5[11] = -44.55 \text{ dBW}$$

User C: The SIR of the *userC* is ‘30’, which is greater than the average SIR. Then $|SIR_{avj}(m) - SIR_k(m)| = 13$; therefore from equation (4) the update coefficient is $x_1 = 0.5$.

The transmit power of userC at the end of the 5th iteration is calculated as,

$$P_k(m + 1) = -57 + 5 \times 0.5[-13] = -72 \text{ dBW}$$

The transmit power of userC at the end of the 10th iteration is calculated as,

$$P_k(m + 1) = -57 + 10 \times 0.5[-13] = -75.13 \text{ dBW}$$

The transmit power of userC at the end of the 16th iteration is calculated as,

$$P_k(m + 1) = -57 + 16 \times 0.5[-13] = -77.17 \text{ dBW}$$

The received power of UserA, UserB and UserC at different iterations is summarized in table 1.

Table 1. The updated transmit powers of mobile users A, B and C at different iterations

Parameter	Mobile UserA	Mobile UserB	Mobile UserC
Initial Power of the mobile user at BS	-60 dBW	-64 dBW	-57 dBW
Updated Power at the end of 5 th iteration	-59.03 dBW	-49.6 dBW	-72 dBW
Updated Power at the end of 10 th iteration	-56.02 dBW	-46.6 dBW	-75.13 dBW
Updated Power at the end of 16 th iteration	-53.98 dBW	-44.55 dBW	-77.17 dBW

The SIR of mobile userA and userB are below the average SIR and far away from the base station as compared to UserC. Therefore received power of userA and userB is increased to -53.98 dBW and -44.55 dBW respectively. Whereas the SIR of mobile userC is above the average SIR and nearer to the base station as compared to user A and userB, therefore the received power is decreased to -77.17 dBW. These power changes should lead to equal SIR for three users in the cell. The Figures 1, 2 and 3 shows the power changes of userA, userB and userC respectively with respect to different power control iterations.

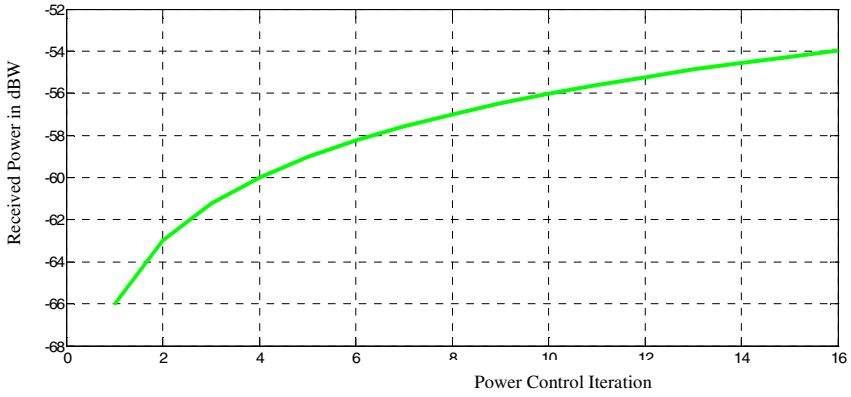


Fig. 1. Updated received Power of UserA

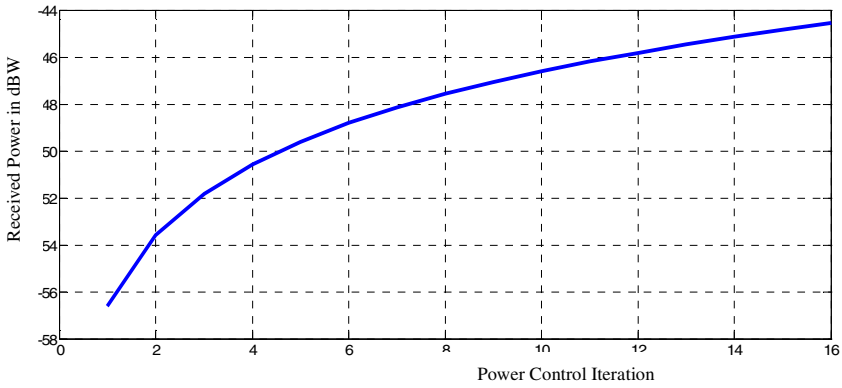


Fig. 2. Updated received Power of UserB

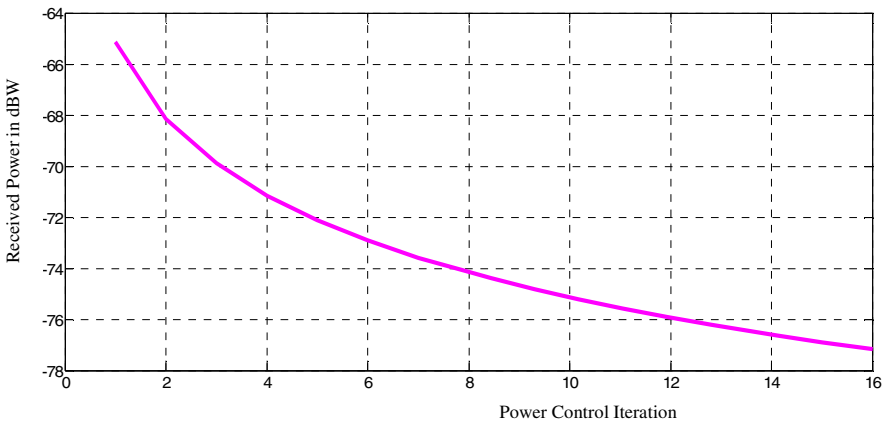


Fig. 3. Updated received Power of UserC

6 Conclusion

The 2.5G network which uses CDMA technology has the advantage of higher capacity and good Quality of Service (QOS) over GSM. Because of non-uniform traffic in CDMA, mobiles users closer to the base station causes a high interference to the mobile users far away from the base station. This causes noise and cancellation of signals in most adversary conditions which leads to decrease in capacity. In this paper a new base station power control algorithm is proposed to improve the user capacity, by optimizing the SIR of the mobile users in network. The proposed algorithm is based on a three-valued power update coefficients (≥ 5 , between 2 and 5 and ≤ 2) with 16 iterations. In this paper three static users A, B and C are assumed in a cell with different distances from the base station. The updated received power of each mobile user is calculated at the end of 5th, 10th and 16th iteration and shown in the figures for 16 iterations. The SIR of mobile userA and userB is considered below the average value, so the power of these mobile users is increased to -53.98 dBW and -44.55 dBW from -60 dBW and -64 dBW; and the SIR of mobile userC is assumed above the average therefore the power of the user is decreased to -77.17 dBW from -57 dBW, to maintain an equal SIR among the users in the cell. Hence the SIR of the three users is optimized to the threshold SIR with the proposed new base station power control algorithm.

References

1. Ziemer, R.E.: Fundamentals of Spread Spectrum Modulation. Morgan & Claypool Publishers, USA (2007)
2. Subramaniam, M., Anpalgan, A.: A brief Introduction to Power Control Techniques in Cellular DS-CDMA Networks. In: Computer Science and Engineering Conference, CSE 2003 (May 2003)
3. Song, W.J., Ahn, B.H., Kim, W.H., Kim, B.G.: Distributed Power Control Based on Bremermann's Genetic Algorithm in CDMA Cellular Radio Networks. In: Proceedings of the 2002 JCCI Annual Conference (2002)
4. Leibnitz, K.: Impacts of Power Control on Outage Probability in CDMA Wireless Networks. In: Proc. IFIP Int.Conf. Broadband Commun., Hong Kong (November 1999)
5. Takeo, K., Sato, S.: The Proposal of CDMA Cell Design Scheme Considering Change in Traffic Distributions. In: Proc. IEEE Int. Symp. Spread Spectrum Tech. and Applic., pp. 229–233 (1998)
6. Hanly, S.: An Algorithm for Combined Cell-Site Selection and Power Control to Maximize Cellular Spread Spectrum Capacity. IEEE Journal on Selected Areas in Communications 13(7), 1332–1340 (1995)
7. Zeeshan, M., Khan, S.A.: Power optimization of multiuser CDMA based mobile ad hoc networks in tactical setting. Proceedings of International Colloquium on Computing, Communication, Control and Management 2, 518–522 (2010)

Performance Evaluation of the Gateway Discovery Approaches under Varying Node Speed

Koushik Majumder¹, Sudhabindu Ray², and Subir Kumar Sarkar²

¹ Department of Computer Science & Engineering,
West Bengal University of Technology, Kolkata, India
koushik@ieee.org

² Department of Electronics and Telecommunication Engineering,
Jadavpur University, Kolkata, India

Abstract. MANETs have emerged as a promising new technology due to their infrastructure-less mode of operation and rapid deployability. In order to facilitate the users with the huge pool of resources together with the global services and applications available from the Internet and for widening the coverage area of the MANET, there is a growing need to integrate the ad hoc networks to the Internet. However, due to the differences in the protocol architecture between MANET and Internet, we need gateways which act as bridges between them. The efficient discovery of gateway in hybrid network is considered as a critical and challenging task due to the scarcity of network resources. With increasing node speed and greater number of sources it becomes even more complex. In this paper, we have conducted a systematic simulation based performance study of the two major gateway discovery approaches using NS2 under different network scenarios. The performance analysis has been done on the basis of three metrics - packet delivery fraction, average end-to-end delay and normalized routing load.

Keywords: Average end-to-end delay, gateway discovery approaches, Internet, Mobile ad hoc network, normalized routing load, packet delivery fraction, performance evaluation.

1 Introduction

With the increasing impact of Internet on our daily life and due to the huge influx of highly portable handheld devices such as smart mobile phones, laptops and personal digital assistants, there is a growing demand for the connectivity to the Internet while we are on the move. In order to provide the connectivity between these devices in the absence of any base station, mobile ad hoc networks (MANET) [1-8] have emerged as a hugely popular solution due to their fast setup capability and non-reliance on any infrastructure or centralized server. But due to the limited transmission range of the MANET nodes, the total area of coverage is often limited. In order to access the huge collection of applications and services from the Internet and to widen the coverage area, there is a growing need to connect these ad hoc networks to the Internet. For this purpose we need Internet Gateways (IGW). There are mainly two types of gateway discovery approaches - proactive gateway discovery [9, 10] and reactive gateway discovery [11, 12].

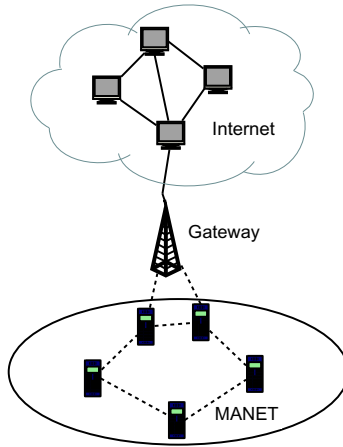


Fig. 1. Hybrid Network

In this work, we have used the extended AODV reactive routing protocol to support the communication between the MANET [13] and the Internet and have studied the performance differentials of the two major gateway discovery approaches – proactive and reactive approaches under increasing node speed with varying number of sources using ns2 based simulation.

The rest of the paper is organized as follows. Section 2 describes the literature study. Section 3 and section 4 details the simulation model and the key performance metrics respectively. The simulation results are presented and analyzed in section 5. Finally the conclusion has been summarized in the section 6. The last section gives the references.

2 Literature Study

There are only few papers available in the literature on the connectivity of MANETs to Internet.

Sun et al. [6.10] investigated the performance of the cooperation between AODV and Mobile IP. This interoperation enables the mobile nodes to connect to the Internet. AODV handles the route discovery and maintenance within the MANET. Mobile IP enables the MANET nodes to connect to the foreign agent and acquire the care of addresses. The foreign agent acts as the Internet Gateway (IGW).

Broch et al. [6.11] presented an initial approach for integrating the MANET with Mobile IP and Internet using Dynamic Source Routing (DSR). In this work the notion of border routers was introduced. Each border router has two interfaces. DSR is used for the interface connected to the MANET, whereas the interface connected to the Internet uses the normal IP routing approach for handling packets coming in and out of the MANET. The nodes which are within the range of the foreign agents act as gateways between Internet and MANET. Foreign agents are discovered following the

reactive approach. They handle the packet forwarding between the MANET and Internet.

Wakikawa et al. [6.12] in their work discussed how a MANET node can derive a globally routable IPv6 address based on the Neighbor Discovery Protocol (NDP) of IPv6. The mobile nodes use this address to connect to the Internet. In this paper two different gateway discovery approaches are designed. The first one is periodic where the gateway floods the gateway advertisement (GWADV) messages at a certain time interval. In the second approach the mobile nodes reactively floods the gateway solicitation (GWSOL) messages.

3 Simulation Model

We have done our simulation based on ns-2.34 [14, 15]. Our main goal was to evaluate the performance of the two major gateway discovery approaches under a range of varying network conditions. The protocols have a send buffer of 64 packets. In order to prevent indefinite waiting for these data packets, the packets are dropped from the buffers when the waiting time exceeds 20 seconds. We have generated the movement scenario files using the *setdest* program which comes with the NS-2 distribution. The total duration of our each simulation run is 800 seconds. We have varied our simulation with movement patterns for six different node speed: 5m/s, 10m/s, 15m/s, 20m/s, 25m/s, 30m/s. In our simulation environment the MANET nodes use constant bit rate (CBR) traffic sources when they send data to the Internet domain. We have used two different communication patterns corresponding to 15 and 25 sources. The complete list of simulation parameters is shown in Table 1.

Table 1. Simulation Parameters

Parameter	Value
Number of Mobile nodes	60
Number of sources	15,25
Number of gateways	2
Number of hosts	2
Transmission range	250 m
Simulation time	800 s
Topology size	1000 m X 600 m
Source type	Constant bit rate
Packet rate	5 packets/sec
Packet size	512 bytes
Pause time	100 seconds
Node speed	5m/s, 10m/s, 15m/s, 20m/s, 25m/s, 30m/s
Mobility model	Random way point
Gateway discovery approaches	Proactive and Reactive

3.1 Hybrid Scenario

We have used a rectangular simulation area of 1000 m x 600 m. The simulation was performed with the first scenario of 60 mobile nodes among which 15 are sources, 2 gateways, 2 routers and 2 hosts and the second scenario of 60 mobile nodes among which 25 are sources, 2 gateways, 2 routers and 2 hosts. For our hybrid network environment we have two gateways located at each side of the simulation area. In our two simulation scenarios, 15 and 25 mobile nodes respectively act as CBR sources.

4 Performance Metrics

We have primarily selected the following three parameters in order to study the performance comparison of the two gateway discovery approaches.

Packet delivery fraction: This is defined as the ratio between the number of delivered packets and those generated by the constant bit rate (CBR) traffic sources.

Average end-to-end delay: This is basically defined as the ratio between the summation of the time difference between the packet received time and the packet sent time and the summation of data packets received by all nodes.

Normalized routing load: This is defined as the number of routing packets transmitted per data packet delivered at the destination.

5 Simulation Results and Analysis

In this section we have analyzed the performance differentials of the proactive and reactive gateway discovery approaches.

5.1 Packet Delivery Fraction (PDF) Comparison

From Figure 2 we observe that, at lower node speed, when the network topology remains relatively stable, the proactive approach shows better packet delivery performance than the reactive approach. At lower node mobility, once the routes are established in the proactive approach, they will remain available for a longer period of time. As all the nodes maintain routes to the gateways all the time, packet delivery can be performed smoothly without having to wait for the path setup time. This results in a better packet delivery performance of the proactive approach. In case of reactive approach, on the contrary, the nodes need not maintain the routes to the gateways all the time. Routes are determined in an on demand basis. When a source node in MANET wants to send a packet to a destination node in the Internet, the source has to first discover the route to the gateway. During this time, no packet can be delivered and the packets waiting at the buffer are ultimately dropped when the maximum buffering time is exceeded. This results in lower packet delivery fraction of the reactive approach in comparison to the proactive approach.

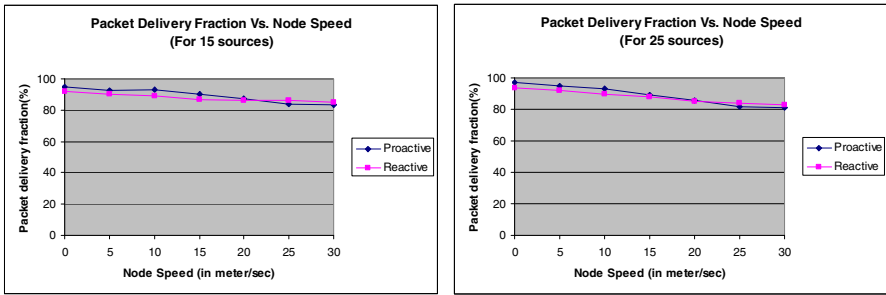


Fig. 2. Packet Delivery Fraction vs. Node Speed for 15 and 25 sources

As the speed of the nodes increases, the network topology becomes highly dynamic and as a result, link breaks become more frequent. Due to the lack of available routes to the gateways, the nodes show deterioration in the packet delivery performance. The proactive approach, due to its periodic nature of operation, becomes less adaptive to this highly dynamic scenario. Once a route breaks, no new route to the gateway can be found until the next gateway discovery interval. This unavailability of routes causes greater number of packets to be dropped in comparison to the reactive approach which is more adaptable to this highly dynamic situation due to its on demand nature of operation.

From the figure it can also be noticed that as the number of sources is increased, initially when the network topology remains relatively stable at lower node speed, the packet delivery fraction also gets better. This happens due to the fact that with lesser number of sources, the channel capacity is not fully utilized. Therefore, increasing the number of sources also increases the packet delivery ratio. However, when the node speed is increased more along with greater number of sources, this leads to congestion and reduced availability of the channel bandwidth for data transmission which ultimately reduces the packet delivery ratio.

5.2 Average End-to-End Delay Comparison

It can be observed from figure 3 that the average end-to-end delay is more with the reactive gateway discovery approach than the proactive gateway discovery approach. In the proactive approach, due to the periodic broadcast of gateway information at fixed intervals, route optimization takes place regularly. As a result, the nodes have access to fresher and shorter routes to the gateway. Moreover, routes to the gateways are maintained at all the nodes all the time. This instant availability of the fresher and shorter routes causes less delay in the delivery of the packets. In reactive approach, on the contrary, routes to the gateways are determined on demand and the gateway discovery precedes the actual delivery of the data packets. This initial path setup delays the delivery of the packets.

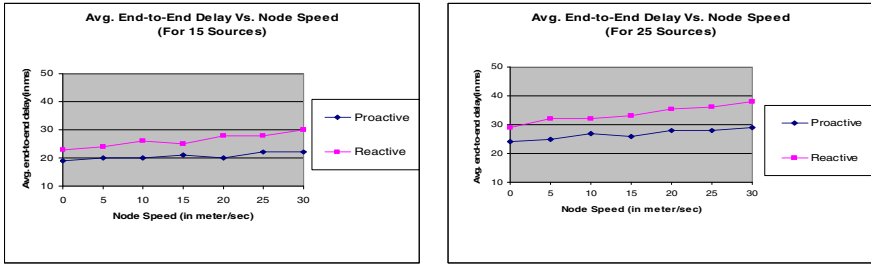


Fig. 3. Average End to End Delay vs. Node Speed for 15 and 25 Sources

From the figure it is evident that the average end-to-end delay increases with greater number of sources and higher node speed for both the approaches. Changes in the network topology become more frequent as the speed of the nodes increases. This results in greater number of link breaks. This together with the greater number of sources necessitates the reactive gateway discovery process to be invoked more frequently in order to find new routes. The frequent invocation of the gateway discovery creates huge amount of control traffic. The data traffic to be delivered also becomes more with greater number of sources. This results in more collisions, further retransmissions and higher congestion in the network. Consequently, the route discovery latency increases due to the constrained channel. This in turn increases the average end-to-end delay. Also the control packets have higher priority over the data packets. Thus the data packets need to spend more time in the queue waiting for the huge volume of control packets to be delivered. This also increases the end-to-end delay in delivering the data packets. In case of proactive approach, due to higher speed of the nodes and frequent link breaks, routes become unavailable and nodes need to wait till the next gateway advertisement for new routes. Thus the delay increases depending upon the duration of the gateway advertisement interval.

5.3 Normalized Routing Load Comparison

From figure 4 we see that the normalized routing load is more in the proactive approach in comparison to the reactive approach. This is primarily due to the periodic broadcast of the gateway advertisement messages to all the mobile nodes in the network irrespective of whether the mobile nodes want them or not. This results in excessive flooding overhead. On the contrary, in the reactive approach, the gateway discovery is initiated in an on demand basis which results in comparatively less routing overhead.

For the proactive approach, the normalized routing load remains almost constant for a particular advertisement interval irrespective of the changes in node speed. Whereas in case of the reactive approach, with increasing node speed, the gateway discoveries need to be invoked more often due to increase in the number of broken links. Furthermore, as the reactive approach does not use route optimization until the route is broken and continues using longer and older routes, the chances of link breaks also increase. This further adds to the number of route discoveries which ultimately results in huge control traffic and subsequently higher normalized routing load.

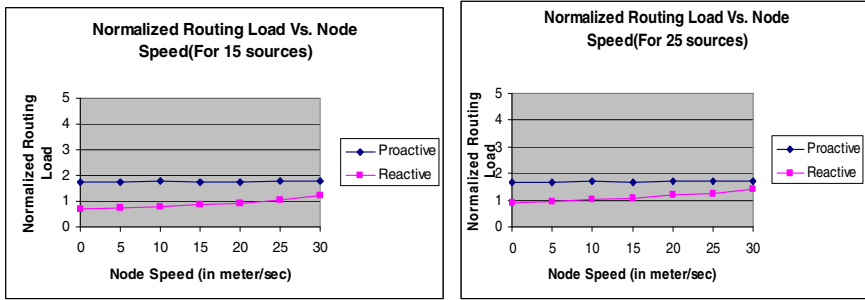


Fig. 4. Normalized Routing Load vs. Node Speed for 15 and 25 Sources

In case of reactive approach, with greater number of sources, the gateway discovery is invoked more often which significantly increases the volume of control overhead. Greater number of sources also creates congestion in the network due to higher volume of data and control traffic. This causes further collisions, more retransmissions and newer route discoveries. This further adds to the already increased control overhead which ultimately results in higher normalized routing load.

From the figure it can be noticed that the normalized routing load is less with more number of sources in case of the proactive approach. This happens because, in the proactive approach, the amount of control overhead remains almost the same for a particular advertisement interval irrespective of the number of sources. But with greater number of sources the number of received data packets is more. This results in the reduced normalized routing load of the proactive approach.

6 Conclusion

In this paper we have carried out a detailed ns2 based simulation to study and analyze the performance differentials of the proactive and reactive gateway discovery approaches under different network scenarios. From the simulation results we see that at lower node speed, the proactive approach shows better packet delivery performance than the reactive approach mainly due to the instant availability of fresher and newer routes to the gateway all the time. On the other hand, with higher node speed, the proactive approach shows more deterioration in the packet delivery performance than the reactive approach mainly due to its less adaptability to the highly dynamic network topology. In terms of the average end-to-end delay, the proactive gateway discovery approach outperforms the reactive gateway discovery. Both the approaches suffer from greater average end-to-end delay when we increase the speed of the nodes and the numbers of sources. As far as normalized routing load is concerned, the reactive approach performs better than the proactive approach. In case of the proactive approach, the amount of control overhead remains almost constant for a particular advertisement interval irrespective of the node speed or the number of sources. With more number of sources, however, the number of received data packets increases for the proactive approach which accounts for its reduced normalized routing load. Whereas for the reactive approach, with faster node speed and greater

number of sources, the number of gateway discoveries and as a result the amount of control traffic also increases, which ultimately results in higher normalized routing load.

References

1. Vaidya, N.H.: Mobile Ad Hoc Networks: Routing, MAC and Transport Issues. In: University of Illinois at Urbana-Champaign, Tutorial Presented at INFOCOM 2004 (IEEE International Conference on Computer Communication) (2004)
2. Arun Kumar, B.R., Reddy, L.C., Hiremath, P.S.: Mobile Ad hoc Networks: Issues, Research Trends and Experiments. *International Engineering and Technology (IETECH) Journal of Communication Techniques* 2(2) (2008)
3. Arun Kumar, B.R., Reddy, L.C., Hiremath, P.S.: A Survey of Mobile Ad hoc Network Routing Protocols. *Journal of Intelligent System Research* (2008)
4. Tanenbaum, A.S.: *Computer Networks*. 4th edn. Prentice Hall
5. Royer, E.M., Toh, C.K.: A Review of Current Routing Protocols for Ad hoc Mobile Wireless Networks. *IEEE Personal Communications Magazine*, 46–55 (1999)
6. <http://www.ietf.org/html.charters/manet-charter.html>
7. Rappaport, T.S.: *Wireless Communications, Principles & Practices*. Prentice Hall (1996)
8. Corson, S., Macker, J.: Mobile Ad hoc Networking (MANET): Routing Protocol Performance Issues and Evaluation Considerations. IETF MANET Working Group RFC-2501 (1999)
9. Jonsson, U., Alriksson, F., Larsson, T., Johansson, P., Maguire Jr., G.Q.: MIPMANET – Mobile IP for Mobile Ad Hoc Networks. In: *The First IEEE/ACM Annual Workshop on Mobile Ad Hoc Networking and Computing (MobiHOC 2000)*, Boston, Massachusetts, USA, pp. 75–85 (2000)
10. Sun, Y., Belding-Royer, E., Perkins, C.: Internet Connectivity for Ad hoc Mobile Networks. *International Journal of Wireless Information Networks*, Special Issue on Mobile Ad Hoc Networks (MANETs): Standards, Research, Applications 9(2) (2002)
11. Broch, J., Maltz, D.A., Johnson, D.B.: Supporting Hierarchy and Heterogeneous Interfaces in Multi-Hop Wireless Ad Hoc Networks. In: *Proceedings of the Workshop on Mobile Computing*, Perth, Australia (1999)
12. Wakikawa, R., Malinen, J.T., Perkins, C.E., Nilsson, A., Tuominen, A.J.: Global connectivity for IPv6 mobile ad hoc networks. draft-wakikawa-MANET-globalv6-03.txt (October 2003)
13. Perkins, C.E.: *Ad hoc networking*. Addison Wesley (2001)
14. Fall, K., Vardhan, K. (eds.): *Ns notes and documentation* (1999), <http://www.mash.cd.berkeley.edu/ns/>
15. Network Simulator-2 (NS2), <http://www.isi.edu/nsnam/ns>
16. IEEE Computer Society LAN MAN Standards Committee. *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, IEEE Std 802.11-1997. The Institute of Electrical and Electronics Engineers, New York (1997)

Speech Enhancement and Recognition of Compressed Speech Signal in Noisy Reverberant Conditions

Maloji Suman^{1,2}, Habibulla Khan², M. Madhavi Latha³,
and Devarakonda Aruna Kumari^{1,2,*}

¹ Life Member, CSI

² K.L University, Vaddeswaram, Guntur, Andhra Pradesh

³ JNTU College of Engineering, Hyderabad

Abstract. Speech compression, enhancement and recognition in noisy, reverberant conditions is a challenging task. In this paper a new approach to this problem, which is developed in the framework of probabilistic random modeling. speech coding techniques are commonly used in low bit rate analysis and synthesis . Coding algorithms seek to minimize the bit rate in the digital representation of a signal without an objectionable loss of signal quality in the process. Speech enhancement aims to improve speech quality by using various algorithms This paper deals with multistage vector quantization technique used for coding of narrow band speech signals. The parameter used for coding of speech signals are the line spectral frequencies, so as to ensure filter stability after quantization. A new approach incorporates the information about statistical random nature of uncompressed speech signal using LBG algorithm .The code books used for quantization are generated by using Linde, Buzo and Gray(LBG) algorithm. Speech model is characterized by LPC coefficients and parameterized by the coefficients of the reverberation filters The results of the multistage vector quantizer are compared with unconstrained vector quantization Technique. The performance of quantization is measured in terms of spectral distortion measured in dB, Computational complexity measured in KFlops and Memory Requirements measured in Floats. From the results it can be proved that multistage vector quantization is having better spectral distortion performance, less computational complexity and memory requirements when compared to unconstrained vector quantization. The proposed approach yields significantly estimating the parameters from the data , better performance in both signal to noise ratio and subjective filter methods

Keywords: Linear predictive Coding, Multi stage vector quantization, Line Spectral Frequencies (LSF).

1 Introduction

The capability of speech compression has been central to the technologies of robust long-distance communication, high-quality speech storage, and message encryption. Compression continues to be a key technology in communications in spite of the promise of optical transmission media of relatively unlimited bandwidth. This is because of our continued and, in fact, increasing need to use band-limited media such

as radio and satellite links, and bit-rate-limited storage media such as CD-ROMs and silicon memories. Storage and archival of large volumes of spoken information makes speech compression essential even in the context of significant increases in the capacity of optical and solid-state memories.

Speech has arguably been most important form of human communication. Since languages were first conceived over the ages, many forms of communication have been developed to convey information across a distance, but the relatively recent invention of the telephone has revalorized this process. The demand for efficient communication & data storage is continuously increasing. One example is the enormous growth in interest communication where image & video signals play an important role. Signal representation is one of the important factor in digital communication. Demand for mobile & convenient forms to communication has been an explosion in the use of cellular & satellite telephony, both of which has significant capacity constraints. The purpose of speech coding research is to address the problem of accommodating more users over such limited capacity by coding speech before transmitting it across a network. As suggested by the research scholar M. satya sai Ram [3][4]

The advantages with coded speech signals are:

- Lower sensitivity to channel noise
- Easier to error-protect, encrypt, multiplex and packetize.
- Efficient transmission over bandwidth constrained channels due to lower bit rate.

The quantization technique should have less computational and memory requirements and it should not result in suboptimal quantization performance intelligibility. Speech coders operating at low bit rates necessitate efficient encoding of linear predictive coding (LPC) coefficients. Line spectral frequencies parameters are currently one of the most efficient choices of transmission parameters for the LPC coefficients.

Multi Stage Vector Quantization can achieve very low encoding and storage complexity in comparison to unstructured vector quantization. However, the conventional MSVQ is suboptimal with respect to the overall performance measure. This paper proposes a new technology to design the decoder codebook, which is different from the encoder codebook to optimize the overall performance. The performance improvement is achieved with no effect on encoding complexity, both storage and time consuming, but a modest increase in storage complexity of decoder. Speech coding is the compression of speech (into a code) for transmission with speech codec's that use audio signal processing and speech processes techniques. The aim of this paper is to provide a general review of MSVQ, and to compare its performance with unconstrained vector quantization technique and this compressed signal is therefore enhanced using enhancement techniques The practical limitations, Regarding computational complexity and memory requirements as a function of bit rate are discussed. spectral distortion performance[6] of MSVQ is evaluated in LSF parameter quantization [7]-[9] for narrow band speech coding. The performance is evaluated by using the spectral distortion method.

Speech enhancement is successful technique with many applications in the filed of speech recognition and speaker recognition.

Speech enhancement depends on environmental conditions and background noise. Reverberation effects are low when background noise level is low and the speaker is close to the microphone. Speech enhancement using spectral subtraction and noise cancelation can enhance the compressed speech signal and offer the satisfactory cancelation of the noise. Spectral subtraction algorithms recover the speech signal of a given training sequence by subtracting the noise spectrum from the signal spectrum, requiring a special treatment when the result is negative. Another example is the difficulty of combining algorithms that remove noise with algorithms that handle reverberation into a single system in a systematic manner. More recently, a new type of speech enhancement algorithms have started to emerge. These algorithms follow by taking a probabilistic modeling approach to the problem. In that approach, one starts by constructing a model for clean speech signals.

2 Dimensions of Performance in Speech Compression

Speech coders attempt to minimize the bit rate for transmission or storage of the signal while maintaining required levels of speech quality, communication delay, and complexity of implementation (power consumption). We will now provide brief descriptions of the above parameters of performance, with particular reference to speech.

Speech Quality

Speech quality is usually evaluated on a five-point scale, known as the mean-opinion score (MOS) scale, in speech quality testing---an average over a large number of speech data, speakers, and listeners. The five points of quality are: bad, poor, fair, good, and excellent. Quality scores of 3.5 or higher generally imply high levels of intelligibility, speaker recognition and naturalness.

Bit Rate

The coding efficiency is expressed in bits per second (bps).

Communication Delay

Speech coders often process speech in blocks and such processing introduces communication delay. Depending on the application, the permissible total delay could be as low as 1 msec, as in network telephony, or as high as 500 msec, as in video telephony. Communication delay is irrelevant for one-way communication, such as in voice mail.

Complexity: The complexity of a coding algorithm is the processing effort required to implement the algorithm, and it is typically measured in terms of arithmetic capability and memory requirement, or equivalently in terms of cost. A large complexity can result in high power consumption in the hardware.

3 Multistage Vector Quantization

Several techniques can be employed in calculating the codebooks in MSVQ design [3][4]. The simplest method is to train the codebooks sequentially. The codebook for the first stage is computed in a traditional manner using, e.g., GLA and the training data is quantized with the obtained one-stage vector quantizer. The resulting quantization error vectors are used as the training data for the second stage. This is repeated for all stages, with each new codebook trained using the error between the original and the reconstructed vectors including all the previous stages.

The multi-stage vector quantizer is a type of product-code vector quantizer which reduces the complexity of a vector quantizer, but at the cost of lower performance. In 2-stage vector quantization [4], the LPC parameter vector (in some suitable representation such as the LSF representation) is quantized by the first-stage vector quantizer and the error vector e (which is the difference between the input and output vectors of the first stage) is quantized by the second-stage vector quantizer. The final quantized version of the LPC vector is obtained by summing the outputs of the two stages. To minimize the complexity of the 2-stage vector quantizer, the bits available for LPC quantization are divided equally between the two stages.

Selection of a proper distortion measure is the most important issue in the design and operation of a vector quantizer. Since the spectral distortion is used here for evaluating LPC quantization performance, ideally it should be used to design the vector quantizer. However, it is very difficult to design a vector quantizer using this distortion measure. Therefore, simpler distance measures (such as the Euclidean and the weighted Euclidean distance measures) between the original and quantized LPC parameter vectors (in some suitable representation such as the LSF representation) are used to design the LPC vector quantizer. To find the best LPC parametric representation for the Euclidean distance measure, the study of the 2-stage vector quantizer with the distance measure in the following three domains: the LSF domain, the arcsine reflection coefficient domain and the log-area ratio domain is done. The 2-stage vector quantizer performs better with the LSF representation than with the other two representations.

The Euclidean distance measure used for vector quantization in the preceding section provides equal weights to individual components of the LSF vector, which obviously are not proportional to their spectral sensitivities. Paliwal and Atal have proposed a weighted Euclidean distance measure in the LSF domain which tries to assign weights to individual LSFs according to their spectral sensitivities. The weighted Euclidean distance measure between the test LSF vector f and the reference LSF vector is given by

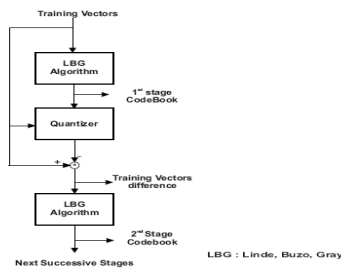


Fig. 1. Codebook generation for different stages of MSVQ

- Initially the codebook at the first stage is generated by using the Linde, Buzo and Gray (LBG) [10] algorithm with the training set as an input.
- Secondly the training difference vectors are extracted by applying the training set and the codebook of the first stage to the quantizer.
- Finally the training difference vectors are used to generate the codebook of the second stage.

4 Complexity and Memory Requirements and Spectral Distortion

Speech enhancement focuses on the suppression of additive background noise as additive noise is easier to deal with than convolutive noise or nonlinear disturbances.

Speech enhancement is a very special case of signal estimation as speech is non-stationary, and the human ear can not judge. Therefore measurements of intelligibility and quality are required.

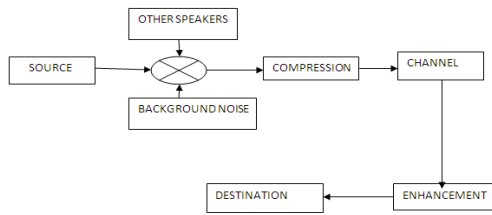


Fig. 2. Speech Enhancement Model

Thus the goal of speech enhancement is to find an *optimal estimate*. It makes use of the fact that power spectra of additive independent signals are also additive and that this property is approximately true for short-time estimates as well. Hence, in the case of stationary noise, it suffices to subtract the mean noise power to obtain a least squares estimate of the power spectrum.

Power spectral subtraction is a minimum mean square estimator with little or no assumptions about the prior distributions for power spectral values of speech and noise. This is the underlying reason why ad hoc operations like clipping are necessary. Within the framework of spectral magnitude estimation two major improvements are: (i) modeling of realistic a priori statistical distributions of speech and noise spectral magnitude coefficients (ii) minimizing the estimation error in a domain which is perceptually more relevant than the power spectral domain (e.g., log magnitude domain) Minimum mean square error estimators (MMSEEs) have been developed under various assumptions such as Gaussian sample distributions, lognormal distribution of spectral magnitudes, etc. While improving on quality, these estimators tend to be complex and computationally demanding.

5 Results

The performance of quantization is measured in terms of Spectral distortion measured in dB, Computational complexity measured in Kflops and Memory Requirements

measured in Floats. Tables 1,2 and 3 shows the spectral distortion(dB), computational complexities (Kflops/frame) and memory requirements (floats) at various bit rates for unconstrained and three stage multistage vector quantizer. From Table-1 it is observed that MSVQ has better spectral distortion performance. From Table-2 & 3 it is observed that MSVQ has less computational complexities and memory requirements when compared to unconstrained vector quantization Technique The code books used for quantization are generated by using Linde, Buzo and Gray(LBG) algorithm. It is observed from Table-1, 2 & 3 that as the number of bits/frame decreases ,the complexity and memory requirements are also decreased but the spectral distortion has increased and transparency in quantization is achieved at 24 bits/frame.

Table 1. Spectral Distortion for MSVQ

Bits / frame	SD(dB)	2-4 dB	>4dB
24(8+8+8)	0.984	1.38	0
23(7+8+8)	1.238	1.2	0.1
22(7+7+8)	1.345	0.85	0.13
21(7+7+7)	1.4	1.08	0.3

Table 2. Complexity, and Memory requirements for unconstrained vector quantization

Bits/Frame	Complexity (k-flops /frame)	ROM(floats)
24	671088.639	167772160
23	335544.319	83886080
22	167772.159	41943040
21	83886.079	20971520
20	41943.039	10485760
19	20971.519	5242880
18	10485.759	2621440

Table 3. Complexity and Memory Requirements for multi stage vector quantization

Bits / frame	Complexity (k-flops/frame)	ROM (floats)
24(8+8+8)	30.717	7680
23(7+8+8)	25.597	6400
22(7+7+8)	20.477	5120
21(7+7+7)	15.357	3840
20(6+7+7)	12.797	3200
19(6+6+7)	10.237	2560
18(6+6+6)	7.677	1920

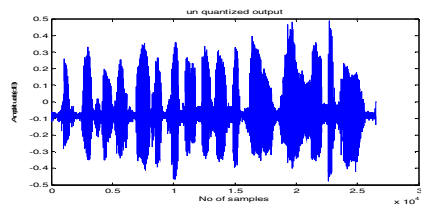


Fig. 3. unconstrained vector quantized speech signal

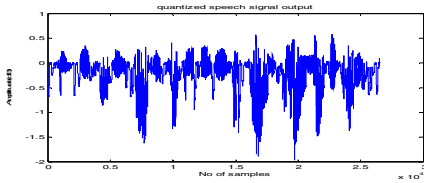


Fig. 4. MSVQ OUTPUT USING 24-BITS

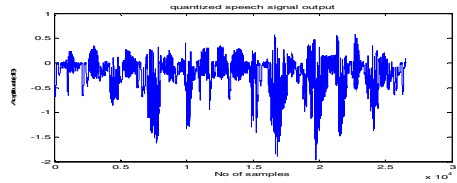


Fig. 5. MSVQ OUTPUT USING 22-BITS

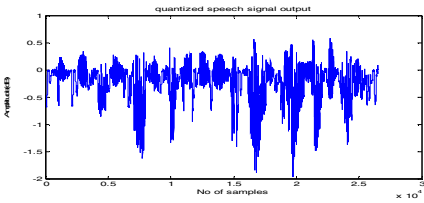


Fig. 6. MSVQ OUTPUT USING 23-BITS

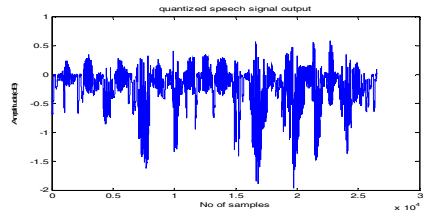


Fig. 7. MSVQ OUTPUT USING 21-BITS

6 Conclusions and Future Work

Speech coding is a method of reducing the amount of information required to represent a speech signal. In this paper two methods of speech coding techniques i.e. unconstrained vector quantization and multi stage vector quantization are analyzed. From results it can be concluded that Multi Stage Vector Quantization is having the less computational complexity & memory requirement when compared to unconstrained vector quantization. But the Spectral distortion performance of the Multi Stage Vector Quantizer is better when compared to unconstrained Vector Quantizer. The decreasing computational complexity & memory requirement with multi stage Vector Quantizer is due to less availability of bits at each stage of quantizer. It may be expected that the use of more complex models, borrowed from speech recognition work, will take us even further. This line of work is promising from a quality point of view but implies much greater computational complexity as well. At the same time these models may have problems in dealing with events that did not occur during the *training phase*.

The proposed research work for speech enhancement in noisy, reverberant situations, based on probabilistic random modeling focused on improvement of SNR and filtering of background noise. Extension this research work can include handling of multi speaker environment and estimation of speech signal in non stationary situations.

References

1. Suman, M., Khan, H., Madhavi Latha, M., Aruna Kumari, D.: Dimensions of performance in compressive speech signals and its enhancement. *International Journal of Engineering Sciences Research (IJSER)* 2(2), 87–93 (2011) ISSN: 2230-8504, e-ISSN : 2230-8512
2. Suman, M., Satya Sai Ram, M., Khan, H.: Compression of Speech signals: LBG Algorithm. In: *Third International Conference on SEEC 2010, Cochin, Kerala* (2010)
3. Madhavi Latha, M., Satya Sai Ram, M., Siddaiah, P.: Multi Switched Split Vector Quantizer. *International Journal of Computer, Information and Systems Science and Engineering* 2(1)
4. Madhavi Latha, M., Satya Sai Ram, M., Siddaiah, P.: Multi Switched Split Vector Quantization. In: *Proceedings of World Academy of Science, Engineering and Technology*, vol. 27 (February 2008) ISSN 1307-6884
5. Harma, A.: Linear predictive coding with modified filter structures. *IEEE Trans. Speech Audio Process.* 9, 769–777 (2001)
6. Stephen, S., Paliwal, K.K.: Efficient product code vector quantization using switched split vector quantiser. *Digital Signal Processing Journal* 17, 138–171 (2007)
7. Paliwal, K.K., Atal, B.S.: Efficient vector quantization of LPCParameters at 24 bits/frame. *IEEE Trans. Speech Audio Process.*, 3–14 (1993)
8. Soong, F., Juang, B.: Line spectrum pair (LSP) and speech data compression. In: *IEEE International Conference on ICASSP*, vol. 9, pp. 37–40 (1984)

Emission Constrained Economic Dispatch Using Logistic Map Adaptive Differential Evolution

Kamal K. Mandal¹, Bidisna Bhattacharya², Bhimsen Tudu¹, and N. Chakravorty¹

¹ Department of Power Engineering, Jadavpur University, Kolkata-700098

² Dept. of Electrical Engineering Techno India, Kolkata

Abstract. A novel adaptive differential evolution based algorithm for solving emission constrained economic dispatch (ECED) problem is presented in this paper. The key factor for successful operation DE is the proper selection of user defined parameters. Choosing suitable values of parameters are difficult for DE, which is usually a problem-dependent task. Unfortunately, there is no fix rule for selection of parameters. The trial-and-error method adopted generally for tuning the parameters in DE requires multiple optimization runs. Even this method can not guarantee optimal results every time and sometimes it may lead to premature convergence. The proposed method combines differential evolution with chaos theory for self adaptation of DE parameters. The performance of the proposed method is demonstrated on a sample test system. The results of the proposed method are compared with other methods. It is found that the results obtained by the proposed method are superior in terms of fuel cost, emission output and losses.

1 Introduction

The main objective of Economic Load Dispatch (ELD) of electric power generation is to schedule the committed generating units so as to meet the load demand at minimum operating cost while satisfying all unit and system equality and inequality constraints. There is a growing need from the society for adequate and secure supply of electricity not only at the cheapest rate, but also at minimum level of emission. Several methods to reduce the atmospheric emissions have been proposed and discussed by many researchers. They include switching to fuels with low emission potential, installing post-combustion cleaning system e.g. electrostatic precipitators, replacement of the aged fuel-burners with cleaner ones, and reallocation of loads to generators with low emission coefficients. The first two methods involve considerable amount of capital investment and hence, can be termed as long term options. The third method is an attractive short- term alternative and requires only minor modification of dispatching programs to include emissions. By proper load allocation among the various generating units of the plants, the harmful effects of the emission of particulate and gaseous pollutants from power stations, particularly from thermal power stations, can be reduced.

Several methods to reduce the atmospheric emissions have been proposed and discussed [1] by many researchers. An earlier attempt on solving minimum emission dispatch problem was taken up by Gent et al. [2]. In recent times, different heuristic

techniques have been applied to solve these complicated problems. Some of these method include back-propagation neural network [3], evolutionary algorithm [4], new recursive technique [5], bio-geography based optimization technique [6] etc.

Differential Evolution (DE) is one of the most recent population-based techniques. DE was originally proposed by Storn and Price in 1995 [7] as a heuristic method for minimizing nonlinear and non differentiable continuous space functions. This paper proposes a novel adaptive differential evolution technique using chaos theory to solve the problem of emission constrained economic dispatch (ECED). The feasibility of the proposed method is demonstrated on a sample test system. The results have been compared with other evolutionary methods and it is found that it can produce comparable results.

2 Problem Description

In this section, we describe the problem formulation for emission constrained economic dispatch (ECED).

2.1 Economic Dispatch

The primary objective of ELD problem is to minimize the total fuel cost of the generating units and to meet the system demand under several operating constraints. Thus, the problem may be described as the minimization of the total fuel cost as defined by (1)

$$FC(P_g) = \sum_{i=1}^n (a_i P_i^2 + b_i P_i + c_i) \tag{1}$$

where, $FC(P_g)$ is total fuel cost of generation in the system (\$/hr), a_i, b_i, c_i are the fuel cost coefficients of the i th generating unit, P_i is the power generated by the i th unit and n is the number of thermal units

The cost is minimized with the following generator capacities and active power balance constraints.

$$P_{i, \min} \leq P_i \leq P_{i, \max} \tag{2}$$

$$\sum_{i=1}^n P_i = P_D + P_L \tag{3}$$

where, $P_{i, \min}$ is the minimum power generation by i th unit, $P_{i, \max}$ is the maximum power generation by i th unit, P_D is the total power demand and P_L is the total transmission loss.

The transmission loss P_L can be calculated by using B matrix and is defined by (4)

$$P_L = \sum_{i=1}^n \sum_{j=1}^n P_i B_{ij} P_j \tag{4}$$

where, B_{ij} 's are the elements of loss coefficient matrix B.

2.2 Emission Dispatch

The emission dispatch problem can be described as the optimization (minimization) of total amount of emission release defined by (5).

$$EC(P_g) = \sum_{i=1}^n (\alpha_i P_i^2 + \beta_i P_i + \gamma_i) \tag{5}$$

where, $EC(P_g)$ is the total amount of emission (lb/hr) and $\alpha_i, \beta_i, \gamma_i$ are the emission coefficients of the i th unit.

2.2 Emission Constrained Economic Dispatch (ECED)

The economic dispatch and emission dispatch are two different problems. Emission dispatch can be included in conventional economic load dispatch problems by the addition of emission cost to the normal dispatch. The bi-objective problem of emission constrained economic dispatch (ECED) can be converted to single objective optimization problem by introducing a price penalty factor h [3] as follows:

$$\text{Minimize } TC = FC(P_g) + h * EC(P_g) \tag{6}$$

where, TC is the total operational cost of the system subject to the constraints defined by (2) and (3).

Now, for a trade off between fuel cost and emission cost (6) can be revised as (7)

Minimize

$$TC = w_1 * FC(P_g) + w_2 * h * EC(P_g) \tag{7}$$

where, w_1 and w_2 are weight factors and (i) $w_1 = 1$ and $w_2 = 0$ for pure economic dispatch (ii) $w_1 = 0$ and $w_2 = 1$ for pure emission dispatch (iii) $w_1 = w_2 = 1$ for emission constrained economic dispatch. The price penalty factor h can be found out by a practical method as discussed by Kulkarni et al [3].

3 Brief Overview of Differential Evolution

DE or Differential Evolution belongs to the class of evolutionary algorithms [7] that include Evolution Strategies (ES) and conventional genetic algorithms (GA). DE is a scheme by which it generates the trial vectors from a set of initial populations. In each step, DE mutates vectors by adding weighted random vector differentials to them. If the fitness of the trial vector is better than that of the target vector, the trial vector replaces the target vector in the next generation.

DE offers several strategies for optimization. The version used here is the DE/rand/1/bin, which is described by the following steps.

3.1 Initialization

The optimization process in DE is carried with four basic operations: initialization, mutation, crossover and selection. The algorithm starts by creating a population

vector P of size N_p composed of individuals that evolve over G generation. Each individual X_i is a vector that contains as many elements as the problem decision variable. Thus,

$$P^{(G)} = [X_1^{(G)}, \dots, X_{N_p}^{(G)}] \tag{9}$$

$$X_i^{(G)} = [X_{1,i}^{(G)}, \dots, X_{D,i}^{(G)}]^T$$

$$i = 1, \dots, N_p \tag{10}$$

The initial population is chosen randomly in order to cover the entire searching region uniformly as follows.

$$X_{j,i}^{(0)} = X_j^{\min} + \sigma_j (X_j^{\max} - X_j^{\min}) \tag{11}$$

where $i = 1, \dots, N_p$ and $j = 1, \dots, D$;

Here D is the number of decision or control variables, X_j^{\min} and X_j^{\max} are the lower and upper limits of the j the decision variables and $\sigma_j \in [0, 1]$ is a uniformly distributed random number generated anew for each value of j . $X_{j,i}^{(0)}$ is the j th parameter of the i th individual of the initial population.

3.2 Mutation Operation

Several strategies of mutation have been introduced in the literature of DE [7]. The mutation operator creates mutant vectors (V_i) by perturbing a randomly selected vector (X_k) with the difference of two other randomly selected vectors (X_l and X_m) according to:

$$V_i^{(G)} = X_k^{(G)} + f_m (X_l^{(G)} - X_m^{(G)}) \tag{12}$$

where X_k , X_l and X_m are randomly chosen vectors $\in [1, \dots, N_p]$ and $k \neq l \neq m \neq i$. The mutation factor f_m that lies within $[0, 2]$ is a user chosen parameter used to control the perturbation size in the mutation operator and to avoid search stagnation.

3.3 Crossover Operation

In order to extend further diversity in the searching process, crossover operation is performed. The crossover operation generates trial vectors (U_i) by mixing the parameter of the mutant vectors with the target vectors according to:

$$U_{j,i}^{(G)} = \begin{cases} V_{j,i}^{(G)} & , \text{ if } \eta_j \leq C_R \text{ or } j = q \\ X_{j,i}^{(G)} & , \text{ otherwise} \end{cases} \tag{13}$$

where $i=1, \dots, N_P$ and $j=1, \dots, D$; η_j is a uniformly distributed random number within $[0, 1]$ generated anew for each value of j . The crossover factor $C_R \in [0, 1]$ is a user chosen parameter that controls the diversity of the population. $X_{j,i}^{(G)}$, $V_{j,i}^{(G)}$ and $U_{j,i}^{(G)}$ are the j th parameter of the i th target vector, mutant vector and trial vector at G generation respectively.

3.4 Selection Operation

Selection is the operation through which better offspring are generated. The evaluation (fitness) function of an offspring is compared to that of its parent. Thus, if f denotes the cost (fitness) function under optimization (minimization), then selection process can be described as

$$X_i^{(G+1)} = \begin{cases} U_i^{(G)}, & \text{if } f(U_i^{(G)}) \leq f(X_i^{(G)}) \\ X_i^{(G)}, & \text{otherwise} \end{cases} \tag{14}$$

The optimization process is repeated for several generations. The iterative process of mutation, crossover and selection on the population will continue until a user-specified stopping criterion, normally, the maximum number of generations allowed, is met.

4 Hybrid Differential Evolution Using Chaos Theory

Optimization algorithms based on chaos theory are stochastic search methodologies and are different from the existing evolutionary algorithms. Evolutionary algorithms use the concepts of bio-inspired genetics and natural evolution. On the other hand, optimization techniques using chaos theory based on ergodicity, stochastic properties, and irregularity. Chaotic sequences display an unpredictable long-term behavior due to their sensitiveness to initial conditions [8]. This feature can be utilized to track the chaotic variable as it travels ergodically over the searching space. This paper utilizes chaotic sequence for automatic adjustment of DE parameters. This helps to escape from local minima and improves global convergence.

One of the simplest dynamic systems evidencing chaotic behavior is the iterator called the logistic map [9] and can be described by the following equation.

$$y(t) = \mu \cdot y(t-1) \cdot [1 - y(t-1)] \tag{15}$$

where t is the sample and μ is control parameter, $0 \leq \mu \leq 4$.

The behavior of the system described by (15) is greatly changed with the variation of μ . The value of μ determines whether y stabilizes at a constant size, oscillates between a limited sequence of sizes, or behaves chaotically in an unpredictable

pattern. Equation (15) is deterministic displaying chaotic dynamics when $\mu = 4$ and $y(0) \notin \{0, 0.25, 0.5, 0.75, 1\}$. In this case, $y(t)$ is distributed in the range of $(0,1)$ provided the initial $y(0) \in (0,1)$.

The values of the parameters mutation factor (f_m) and cross over ratio (C_R) can be modified using (16) and (17) as follows:

$$f_m(G) = \mu \cdot f_m(G-1) \cdot [1 - f_m(G-1)] \tag{16}$$

$$C_R(G) = \mu \cdot C_R(G-1) \cdot [1 - C_R(G-1)] \tag{17}$$

where G is the current iteration number.

5 Results and Discussions

The proposed method has been applied to a six-generator test system to verify its effectiveness.

The cost coefficients, generation limits and emission coefficients, valve point coefficients are derived from [6], [10]. The population size $N_p = 60$, the initial mutation factor $f_m = 0.55$ and the crossover factor $C_R = 0.80$ are considered for the study. Maximum iteration number is set at 300.

Table 1. Solution for Six-Generator System – Economic Dispatch and Emission Dispatch

Unit (MW)	Economic Dispatch		Emission Dispatch	
	Demand (MW)		Demand(MW)	
	500	900	500	900
P1 (MW)	35.50	125.00	37.58	125.00
P2 (MW)	15.00	62.01	55.47	113.98
P3 (MW)	80.50	75.71	84.36	126.71
P4 (MW)	90.30	140.73	68.78	152.36
P5 (MW)	115.00	322.05	130.39	249.15
P6 (MW)	179.70	220.32	141.27	193.21
Total Generation (MW)	516.50	945.83	517.85	960.42
Losses (MW)	16.50	45.83	17.85	60.41
CPU Time (Sec).	4.78	23.82	4.65	35.29
Iterations	300	300	300	300
Fuel Cost (Rs./hr)	27890.00	49493.00	28338.00	51010.00
Emission Output (Kg/hr)	282.78	817.93	272.23	740.39

The problem is first solved as a pure economic dispatch problem with and then pure emission dispatch problem. Results are shown in Table 1. Finally emission constrained economic dispatch (ECED) problem and results are shown in Table 2 for the demand of 500 MW and 900 MW. The randomness of the proposed method has been verified by testing with same demand for several times.

Table 2. Solution For Six-Generator system – Emission constrained Economic Dispatch (ECED)

Unit (MW)	Demand (MW)	
	500	900
P1 (MW)	46.82	100.00
P2 (MW)	35.75	98.76
P3 (MW)	94.46	123.41
P4 (MW)	53.32	128.56
P5 (MW)	140.39	269.14
P6 (MW)	146.27	231.34
Total Generation (MW)	517.01	951.21
Losses (MW)	17.01	51.21
CPU Time (Sec).	4.81	37.61
Iterations	300	300
Fuel Cost (Rs./hr)	28149.00	49532.00
Emission Output (Kg/hr)	279.55	746.21

The success rate of the proposed method in finding the global solution is found to be almost 100%. It is noted that CPU time is almost same for a particular demand for all the cases of ELD, ELD and ECED.

Table 3. Comparison of results by different methods for ECED

Demand (MW)	Methods	Fuel Cost (Rs./hr)	Emission (Kg/hr)
500	NR [10]	28550.15	312.513
	FCGA [10]	28231.06	304.90
	NSGA [10]	28291.11	284.362
	BBO [6]	28318.50	279.30
	Proposed Method	28149.00	279.55
900	NR [10]	50807.24	864.06
	FCGA [10]	49674.28	850.29
	NSGA [10]	50126.05	784.69
	BBO [6]	50297.27	765.08
	Proposed Method	49532.00	746.21

The performance of the proposed method is compared with Newton-Raphson, fuzzy controlled genetic algorithm (FCGA) method [10], NSGA and BBO method [6]. The results are shown in Table 3. It is observed that proposed DE based method can provide better results compared with other classical as well as modern population based heuristic methods in terms of fuel cost and emission release.

6 Conclusion

Environmental concern is one of the important issues in the operation of present day power systems. In this paper we have successfully implemented a novel adaptive DE based optimization technique to solve emission constrained economic dispatch ECED problems with equality as well as non-equality constraints. The results of the proposed method based on DE are compared with fuzzy controlled genetic algorithm (FCGA), Newton-Raphson method and bio-geography based optimization technique

for six-generator system. It has been seen that proposed method provides better result in terms of fuel cost, emission output and losses. It is also noted that computation time is considerably reduced in comparison with other methods.

Acknowledgments. We would like to acknowledge and thank Jadavpur University, Kolkata, India for providing all the necessary help to carry out this work.

References

- [1] Talaq, J.H., El-Hawary, F., El-Hawary, M.E.: A summary of environmental/economic dispatch algorithms. *IEEE Transaction on Power Systems* 9, 1508–1516 (1994)
- [2] Gent, M.R., Lamont, J.W.: Minimum Emission Dispatch. *IEEE Transaction on Power Systems PAS-90*, 2650–2660 (1972)
- [3] Kulkarni, P.S., Kothari, A.G., Kothari, D.P.: Combined economic and emission dispatch using improved backpropagation neural network. *Electric Power Components and System* 28, 31–44 (2000)
- [4] Abido, M.A.: Environmental/Economic Power Dispatch Using Multiobjective Evolutionary Algorithms. *IEEE Transactions on Power Systems* 18(4), 1529–1537 (2003)
- [5] Muralidharan, S., Subramanian, S., Srikrishna, K.: Emission constrained economic dispatch – A new recursive approach. *Electric Power Components and Systems* 34, 343–353 (2006)
- [6] Roy, P.K., Ghoshal, S.P., Thakur, S.S.: Combined Economic Emission dispatch biogeography based optimization. *Electrical Engineering* 92, 173–184
- [7] Storn, R., Price, K.: Differential evolution: A simple and efficient adaptive scheme for global optimization over continuous spaces. Technical Report TR-95-012. International Computer Science Institute, Berkeley (1995)
- [8] Yuan, X., Cao, B., Yang, B., Yaun, Y.: Hydrothermal scheduling using chaotic hybrid differential evolution. *Energy Conversion and Management* 49, 3627–3633 (2008)
- [9] May, R.: Simple mathematical models with very complicated dynamics. *Nature* 261, 459–467 (1976)
- [10] Song, Y.H., Wang, G.S., Wang, P.Y., Johns, A.T.: Environmental /Economic Dispatch using Fuzzy Logic Controlled Genetic Algorithms. *IEE Proc. on Generation, Transmission, Distribution* 144(4), 377 (1997)

PSO Based Edge Keeping Suppression of Impulses in Digital Imagery

Jyotsna Kumar Mandal and Somnath Mukhopadhyay

Department of Computer Science and Engineering,
University of Kalyani, Kalyani,
West Bengal, India, 741235
jkm.cse@gmail.com, som.cse@live.com

Abstract. This paper proposes an efficient switching median filter to restore digital images corrupted by high density of random valued impulse noises. The noise detection is performed using all neighbor directional weighted pixels in the 5 x 5 window. Arithmetic absolute differences and intensity of the center pixel is compared with other pixels in the test window to define a noisy pixel. To restore the noisy pixel variable window based median filtering has been done. Particle swarm optimization(PSO), a recent stochastic global optimization technique has been adopted to obtain best fitted parameters of the proposed detection and filtering operators. Simulation results, conducted on a variety of gray scale images clearly exhibit that the proposed operator obtains better results compared to existing directional weighted filters.

1 Introduction

Digital image gets corrupted by impulses during acquisition or transmission because of perturbation in sensors and communication channels. Most common types of impulse are salt-and-pepper(SPN) [10] and random-valued-impulse-noise(RVIN). A linear filter degrades the image seriously as it smoothes the image even with low noise density [10]. The non linear filters are most popular than the linear such as median filtering technique. The standard median (SM) filter [3] provides a standard noise removal performance but removes image fine textures even at low noise ratios. The weighted median (WM) filter, center weighted median (CWM) filter [11] and adaptive center weighted median (ACWM) filter [5] are improved median filters. Hence several switching median filters have been proposed in the literature, which use an impulse detector prior to filter the noises, such as, an iterative pixel-wise modification of MAD (PWMAD) (median of the absolute deviations from the median) filter [7], tri-state median (TSM) filter [4], multi-state median (MSM) filter [6], progressive switching median filter (PSM) [17] and the signal-dependent rank ordered mean filter (SD-ROM) [2]. Some advanced switching median filters such as directional weighted median filter [8], second order difference based impulse detection filter [16], MWB [13] and MDWMF [14] have been proposed in the literature to remove RVIN in the digital images. Several soft computing tools based

filters have also been proposed in the literature such as fuzzy filter [15], neuro fuzzy filter [12], etc to remove impulses in the images.

In this paper, we have proposed all neighbor directional weighted pixels based median filter and corresponding user parameters of the algorithms viz., number of Iterations (I), Threshold (T) and decreasing Rate (R) of threshold in each iteration are searched in a 3-dimensional space to obtain global optimal solution using a stochastic search strategy, i.e., particle swarm optimization (PSO) technique.

The rest of the paper organized as follows. The proposed impulse detection and filtering method are given in sections 2 and 3 respectively. PSO based algorithm is described in section 4. Experimental results and discussions are fabricated in section 5. Conclusions are given in section 6.

2 Impulse Detector

The proposed noise detection scheme given in algorithm 1 is applied on each 5 x 5 window of the image in row major order to classify center pixel which emphasizes on the pixels aligned in the four main directions along with two end pixels in each direction, shown in fig 1. The directional pixels are treated as edges within the test window so that the edges are detected and preserved.

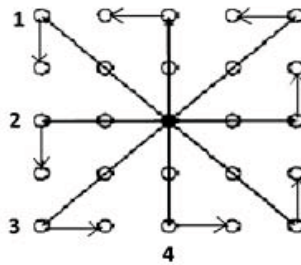


Fig. 1. All neighbor directional weighted pixels in the 5 x 5 window

Algorithm 1. Impulse detector

1: The center pixel $y_{i,j}$ is classified by finding the minimum and maximum gray values of the 5 x 5 window around it. If the value of $y_{i,j}$ does not lie within the range spread of its neighbors, it is declared as noisy. Otherwise we assume that it may not be noisy pixel and considered for next level of detection. Let W_{min} and W_{max} are the maximum and minimum intensity values respectively within the window, then the first level detection rule is given as:

$$y_{i,j} = \begin{cases} \text{Undetected} & : W_{min} < y_{i,j} < W_{max} \\ \text{Noisy} & : W_{min} \geq y_{i,j} \geq W_{max} \end{cases} \quad (1)$$

2: Let the set of seven pixels centered at (0, 0) in the k^{th} direction is S_k ($k=1$ to 4), i.e.,

- . $S_1 = \{(-1,-2), (-2,-2), (-1,-1), (0,0), (1,1), (2,2), (1,2)\}$.
- . $S_2 = \{(1,-2), (0,-2), (0,-1), (0,0), (0,1), (0,2), (-1,2)\}$.
- . $S_3 = \{(2,-1), (2,-2), (1,-1), (0,0), (-1,1), (-2,2), (-2,1)\}$.
- . $S_4 = \{(-2,-1), (-2,0), (-1,0), (0,0), (1,0), (2,0), (2,1)\}$.

Then let $S_k^0 = S_k / (0,0)$, $\forall k = 1$ to 4.

3: In each direction of a 5 x 5 window, define $d_{i,j}^{(k)}$ as the sum of absolute differences of intensity values between $y_{i+s,j+t}$ and $y_{i,j}$ with $(s,t) \in S_k^0$ ($k= 1$ to 4), given in eqn 2. The center of the test window is (i, j).

4: In each direction, weigh the absolute differences between two closest pixels from the center pixel with a large ω_m , between the center pixel and the corner pixels by ω_n and between two end pixels from the center pixel with a small ω_o , before calculating the sum. Assign $\omega_m = 2$, $\omega_n = 1$ and $\omega_o = 0.5$.

Define $d_{i,j}^{(k)}$ as $(\sum_{(s,t) \in S_k^0} \omega_{s,t} |y_{i+s,j+t} - y_{i,j}|, 1 \leq k \leq 4)$, where (2)

$$\omega_{s,t} = \begin{cases} \omega_m & : (s,t) \in \Omega^3 \\ \omega_o & : (s,t) \in \Omega^2 \\ \omega_n & : \text{otherwise} \end{cases} \quad (3)$$

where $\Omega^3 = \{(s,t) : -1 \leq s, t \leq 1\}$, and (4)

where $\Omega^2 = \{(s,t) : (s,t) = \pm\{(-2,-1), (-1,-2), (1,-2), (2,-1)\}\}$. (5)

5: $d_{i,j}^{(k)}$ is known as direction index. Calculate the minimum of these four direction indices for impulse detection, which is given by

$$r_{i,j} = \min\{d_{i,j}^{(k)} : 1 \leq k \leq 4\} \quad (6)$$

Three assumptions are deployed depending upon the values $r_{i,j}$.

- 1. $r_{i,j}$ is small when $y_{i,j}$ is on a noise free flat region.
- 2. $r_{i,j}$ is small when $y_{i,j}$ is on the edge.
- 3. $r_{i,j}$ is large when $y_{i,j}$ is noisy .

6: Form the complete decision rule to detect a noisy or noise free pixel depending upon the definition of $r_{i,j}$, by introducing a threshold (T) and which is given as

$$y_{i,j} = \begin{cases} \text{Noisy Pixel} & : W_{min} \geq y_{i,j} \geq W_{max} \\ \text{Noise Free Pixel} & : r_{i,j} \leq T \text{ and } W_{min} < y_{i,j} < W_{max} \end{cases} \quad (7)$$

3 Impulse Filter

If any pixel is detected as noisy, the filtering scheme prescribed in algorithm 2 restores it to a pixel which is most suitable in the 5 x 5 window.

Algorithm 2. Impulse filter

1: Calculate the standard deviations $\sigma_{i,j}^{(k)}$ of all $y_{i+s,j+t}$ with $(s,t) \in S_k^0$ ($k=1$ to 4).

2: Find the minimum of $\sigma_{i,j}^{(k)}$, where $k=1$ to 4, as

$$l_{i,j} = \min_k \{ \sigma_{i,j}^{(k)} : k = 1 \text{ to } 4 \} \quad (8)$$

3: Find the maximum of $\sigma_{i,j}^{(k)}$, where $k=1$ to 4, as

$$m_{i,j} = \max_k \{ \sigma_{i,j}^{(k)} : k = 1 \text{ to } 4 \} \quad (9)$$

4: Select the directions where standard deviations are maximum, minimum and also not any of them. Use repetition operator \diamond [3] for which standard deviation is minimum.

5: Calculate the median using eqn. 10 with assigning $\omega_m = 0$, $\omega_l = 2$ and $\omega_n = 1$.

$$med = median\{ \omega_{s,t} \diamond y_{i+s,j+t} : (s,t) \in \Omega^4 \}, \text{ where} \quad (10)$$

$$\Omega^4 = \{ (s,t) : -1 \leq s, t \leq 1 \} \text{ and } (s,t) \neq (0,0), \text{ and where} \quad (11)$$

$$\omega_{s,t} = \begin{cases} \omega_m & : (s,t) \in S_{m_{i,j}}^0 \\ \omega_l & : (s,t) \in S_{l_{i,j}}^0 \\ \omega_n & : \text{otherwise} \end{cases} \quad (12)$$

6: Replace $y_{i,j}$ by med .

4 Optimization Using PSO

Three user parameters viz., I(maximum number of iterations), T(threshold) and R(decreasing rate of threshold in each iteration) are searched in a 3-D space to obtain optimal solutions using a relatively recent stochastic global optimization technique i.e., particle swarm optimization (PSO). It is chosen because of its fast convergence rate and ease of implementation. It is a population based search and optimization technique developed by Dr. Eberhart and Dr. Kennedy in 1995 [1]. The proposed PSO model has been presented in algorithm 3.

4.1 Performance Metric

The aim is to maximize the value of PSNR using eqn. 15 as a result this equation is used as fitness f for the particles in PSO based optimization technique. Here M and

N are the dimensions of the input images respectively. I_1 and I_2 are the original and enhanced images respectively.

$$f(I_1, I_2) = PSNR(dB) = 10 * \log_{10} \left(\frac{255^2}{\frac{1}{M*N} \sum_{m,n} (I_1m,n - I_2m,n)^2} \right) \tag{15}$$

Algorithm 3. PSO based optimization algorithm

- 1: Three dimensional search space represented by parameters I, T and R are initialized in [3, 8], [300, 999] and [0.6, 0.95] respectively. They are initialized randomly at x_p position in a fixed size of swarm. 'p' represents particle number in a swarm. At the initial position x_p , fitness values f_p are calculated using eqn. [15](#)
- 2: The new positions $x_p(i+1)$ of the particles are searched after calculating the velocities $v_p(i+1)$ using eqns. [13](#) and [14](#) respectively.

$$v_p(i+1) = h(i) * v_p(i) + \Psi_p * r_p * (x_{pbb}(i) - x_p(i)) + \Psi_g * r_g * ((x_{gbb}(i) - x_p(i)) \tag{13}$$

$$x_p(i+1) = x_p(i) + v_p(i+1) \tag{14}$$

- 3: Ψ_p and Ψ_g are the positive learning factors respectively. We have used Ψ_p and $\Psi_g > 1$. r_p and r_g are random numbers in [0, 1], generated in every generation separately. i is the generation number initialized to 1 and I_{MAX} is the maximum number of generations, i.e., [15, 20]. $h(i)$ is the inertia factor, which has positive real random values in less than 1. $x_p(i)$ and $v_p(i)$ are position and velocity of the p^{th} particle at i^{th} iteration, respectively. $f_{pB}(i)$ and $f_{gB}(i)$ are the pBest (personal best fitness value of a particle) value and gBest (global best fitness value of particles) values at i^{th} generation, respectively. $x_{pB}(i)$ and $x_{gB}(i)$ are the personal best positions and the global best position of p^{th} particle at i^{th} generation respectively. These values are initialized by assigning locations of particles where $f_{pB}(i)$ and $f_{gB}(i)$ have been obtained respectively.
- 4: The velocities and positions of particles are updated using eqns. [13](#) and [14](#) respectively.
- 5: To keep the new velocities in the search boundary, the boundary values have been set in $[v_{min}, v_{max}]$. If new velocities and new positions of the particles are found beyond the boundaries of velocities and search space then they are restricted to the boundary values of the space by taking random velocities and positions again for that particular particle.
- 6: At each new position $x_p(i+1)$ of the particles, the fitness values $f_p(i+1)$ are calculated using eqn. [15](#)
- 7: The $f_p(i+1)$ calculated in step 5 is compared with its previous $f_{pB}(i)$. If $f_p(i+1)$ is better than previous $f_{pB}(i)$ then $f_{pB}(i+1)$ is updated by $f_p(i+1)$, otherwise old $f_{pB}(i)$ is retained as a current $f_{pB}(i+1)$. Similarly $x_{pB}(i+1)$ is also updated according to this updated fitness $f_{pB}(i+1)$.
- 8: Best value among the all current $f_{pB}(i+1)$ calculated in step [7](#) is considered as new $f_{gB}(i+1)$. If new value of $f_{gB}(i+1)$ is better than previous $f_{gB}(i)$ then values of $f_{gB}(i)$ is updated by new $f_{gB}(i+1)$, otherwise old $f_{gB}(i)$ is retained as new $f_{gB}(i+1)$. Similarly, $x_{gB}(i+1)$ is also updated according to this updated fitness $f_{gB}(i+1)$.
- 9: Steps [4](#) to [8](#) are repeated until an adequate fitness is reached or a desired maximum number of iterations are attained, but for present implementation the interval [15, 20] is taken as maximum number for iteration.

5 Results and Discussions

The performance of the proposed operator is implemented under various noise densities and on several popular 8 bit gray scale images with dimensions of 512 x 512 like *Boats*, *Bridge* and *Lena* etc,. The algorithm have been executed on the machine configuration as ACPI uni-processor with Intel® Pentium® E2180 @ 2.00 Ghz CPU and 2.98 Gbyte RAM with MATLAB 8a environment.

Fig. 2 shows the comparative visual restoration effects between the existing algorithms and the proposed filter when the *Lena* image is 60% noisy. Considering very high noise ratio and fine details/textures of the images, the proposed filter can enhance the noisy image effectively. The restoration results of the proposed algorithm are compared with existing techniques on *Lena*, *Bridge* and *Boat* images corrupted with 40% to 60% noise densities and given in table 1, 2 and *Bridge* respectively. It is seen from these tables, the performances of proposed operator is best compared to existing methods taken into consideration in restoring the noise densities considered.

Table 1. Comparison of restoration results in terms of *PSNR (dB)* for *Lena* image

Filter	40% Noisy	50% Noisy	60% Noisy
SM[3]	27.64	24.28	21.58
PSM[17]	28.92	26.12	22.06
ACWM[5]	28.79	25.19	21.19
MSM[6]	29.26	26.11	22.14
SD-ROM[2]	29.85	26.80	23.41
Iterative Median[9]	30.25	24.76	22.96
Second Order[16]	30.90	28.22	24.84
PWMAD[7]	31.41	28.50	24.30
DWM Filter[8]	32.62	30.26	26.74
Proposed	32.95	30.92	28.62

Table 2. Comparison of restoration results in terms of *PSNR (dB)* for *Bridge* image

Filter	40% Noisy	50% Noisy	60% Noisy
ACWM[5]	23.23	21.32	19.17
MSM[6]	23.55	22.03	20.07
SD-ROM[2]	23.80	22.42	20.66
Second Order[16]	23.73	22.14	20.04
PWMAD[7]	23.83	22.20	20.83
DWM Filter[8]	24.28	23.04	21.56
Proposed	24.79	24.42	23.95



Fig. 2. Results of different filters in restoring 60% corrupted image Lena, (a) Original image (b) Noisy Image (c) (SD-ROM) [2] (d) (MSM) [6] (e) (PWMAD) [7] (f) (DWM) [8] (g) **Proposed**

Table 3. Comparison of restoration results in terms of *PSNR (dB)* for *Boat* image

Filter	40% Noisy	50% Noisy	60% Noisy
ACWM [5]	26.17	23.92	21.37
MSM [6]	25.56	24.27	22.21
SD-ROM [2]	26.45	24.83	22.59
PWMAD [7]	26.56	24.85	22.32
DWM Filter [8]	27.03	25.75	24.01
Proposed	28.25	27.90	26.41

6 Conclusion

This paper proposed a novel algorithm with PSO based optimization for suppressing the digital images corrupted with random valued impulses. Four directions of the test window are considered as edges within the window, as a result the algorithm preserve edges during noise suppression. The detection operator computes simple arithmetic comparisons and operations on the pixels in the window. The noisy pixel is replaced by median value of some particular pixels of the 3 x 3 window to restore it. Three user parameters of the proposed algorithm are searched in the 3D space using a stochastic global optimization technique, PSO. Better performance is obtained using the proposed operator in terms of subjective quality in restored image and objective quality in terms of *PSNR* (*dB*) compared to existing algorithms.

References

1. Swarmintelligence, <http://www.swarmintelligence.org/> (accessed on August 25, 2011)
2. Abreu, E., Lightstone, M., Mitra, S.K., Arakawa, K.: A new efficient approach for the removal of impulse noise from highly corrupted images. *IEEE Transactions on Image Processing* 5(6), 1012–1025 (1996)
3. Brownrigg, D.R.K.: The weighted median filter. *Communications of the ACM* 27(8), 807–818 (1984)
4. Chen, T., Ma, K., Chen, L.: Tri- state median filter for image de noising. *IEEE Transaction Image Processing* 8(12), 1834–1838 (1999)
5. Chen, T., Wu, H.R.: Adaptive impulse detection using center weighted median filters. *IEEE Signal Processing Letters* 8(1), 1–3 (2001)
6. Chen, T., Wu, H.R.: Space variant median filters for the restoration of impulse noise corrupted images. *IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing* 48(8), 784–789 (2001)
7. Crnojevic, V., Senk, V., Trpovski, Z.: Advanced impulse detection based on pixel- wise mad. *IEEE Signal Processing Letters* 11(7), 589–592 (2004)
8. Dong, Y., Xu, S.: A new directional weighted median filter for removal of random - valued impulse noise. *IEEE Signal Processing Letters* 14(3), 193–196 (2007)
9. Forouzan, A.R., Araabi, B.: Iterative median filtering for restoration of images with impulsive noise. *Electronics, Circuits and Systems* 1, 232–235 (2003)
10. Gonzalez, R.C.: *Woods: Digital image processing*, 2nd edn. Pearson Prentice-Hall (2002)
11. Ko, S.J., Lee, Y.H.: Center weighted median filters and their applications to image enhancement. *IEEE Transactions on Circuits and Systems* 38(9), 984–993 (2001)
12. Kong, H., Guan, L.: A neural network adaptive filter for the removal of impulse noise in digital images. *Neural Networks Letters* 9(3), 373–378 (1996)
13. Mandal, J.K., Sarkar, A.: A novel modified directional weighted median based filter for removal of random valued impulse noise. In: *International Symposium on Electronic System Design (ISED)*, pp. 230–234 (December 2010)
14. Mandal, J.K., Sarkar, A.: A modified weighted based filter for removal of random impulse noise. In: *Second International Conference on Emerging Applications of Information Technology*, pp. 173–176 (February 2011)
15. Russo, F., Ramponi, G.: A fuzzy filter for images corrupted by impulse noise. *IEEE Signal Processing Letter* 3, 168–170 (1996)

16. Sa, P.K., Dash, R., Majhi, B.: Second order difference based detection and directional weighted median filter for removal of random valued impulsive noise. *IEEE Signal Processing Letters*, 362–364 (December 2009)
17. Wang, Z., Zhang, D.: Progressive switching median filter for the removal of impulse noise from highly corrupted images. *IEEE Transactions on Circuits and Systems* 46(1), 78–80 (1999)

Non-recursive FIR Band Pass Filter Optimization by Improved Particle Swarm Optimization

Sangeeta Mandal¹, Sakthi Prasad Ghoshal¹, Rajib Kar², Durbadal Mandal²,
and S. Chaitnya Shiva²

¹Department of Electrical Engg.

National Institute of Technology, Durgapur, West Bengal, India

²Department of Electronics and Communication Engineering

National Institute of Technology, Durgapur, West Bengal, India

rajibkarece@gmail.com

Abstract. This paper proposes a novel optimal design of linear phase digital band pass finite impulse response (FIR) filter using Improved Particle Swarm Optimization (IPSO) technique. IPSO is an improved particle swarm optimization (PSO) that proposes a new definition for the velocity vector and swarm updating and hence the solution quality is improved. Evolutionary algorithms like real code genetic algorithm (RGA), PSO, IPSO have been used here for the design of linear phase band pass FIR filter. A comparison of simulation results reveals the optimization efficacy of the algorithm over the prevailing optimization techniques for the solution of the multimodal, non-differentiable, highly non-linear, and constrained filter design problems.

1 Introduction

Digital signal processing (DSP) presents greater flexibility, higher performance (in terms of attenuation and selectivity), better time and environment stability along with lower equipment production costs than traditional analog techniques. A digital filter is simply a discrete-time, discrete-amplitude convolver. A filter is designed with a frequency domain impulse response which is as close to the desired ideal response as can be generated given the constraints of the implementation.

Traditionally, different techniques exist for the design of digital filters. Out of these, windowing method is the most popular. Evolutionary methods have been employed in the design of digital filters to design with better parameter control and to better approximate the ideal filter. Different heuristic optimization algorithms such as genetic algorithm (GA) [1-3], simulated annealing algorithms [4], Tabu search [5], and artificial bee colony algorithm [6], PSO [8] have been widely used to design optimal digital filters.

The approach detailed in this paper takes advantage of the power of the stochastic global optimization technique called improved particle swarm optimization. PSO is an evolutionary algorithm developed by Eberhart *et al.* [7]. The PSO is simple to implement and its convergence may be controlled via few parameters. The limitations of the conventional PSO are that it may be influenced by premature convergence and stagnation problem [9]. In order to overcome these problems, the PSO algorithm has been modified in this paper and is employed for FIR band pass filter design.

This paper describes an alternative approach for the FIR band pass digital filter design using Improved Particle Swarm Optimization Approach (IPSO). IPSO algorithm tries to find the best coefficients that closely match the ideal frequency response. Based upon the IPSO approach, this paper presents a good and comprehensive set of results, and states arguments for the superiority of the algorithm. Simulation result demonstrates the effectiveness and better performance of the proposed designed method.

2 Band Pass FIR Filter Design

A digital FIR filter is characterized by,

$$H(z) = \sum_{n=0}^N h(n)z^{-n}, \quad (1)$$

where N is the order of the filter which has $(N+1)$ number of coefficients $h(n)$ as the filter impulse responses. The values of $h(n)$ will determine the type of the filter e.g. low pass, high pass, band pass etc. and are to be determined in the design process. This paper presents the most widely used FIR filter with $h(n)$ as even symmetric and the order is even. The length of $h(n)$ is $N+1$ and the number of coefficients is also $N+1$. Because the coefficients $h(n)$ are symmetrical, the dimension of the problem is halved. The $(N+1)/2$ coefficients are then flipped and concatenated to find the required $(N+1)$ number of coefficients. The optimization algorithm attains the minimum error between the desired frequency response and the actual frequency response by determining the optimal $h(n)$ values after a certain maximum number of iterations. The optimal $h(n)$ values, after concatenation, finally represent the filter with better frequency response.

Various filter parameters which are responsible for the optimal filter design are stop band and pass band normalized edge frequencies (ω_p, ω_s), pass band and stop band ripples (δ_p and δ_s), stop band attenuation and transition width. These parameters are mainly decided by the filter coefficients. In this paper, IPSO is applied in order to obtain the desired filter response as close as possible to the ideal response, where $\delta_p, \delta_s, N, \omega_p, \omega_s$ are individually specified.

Now for (1), each filter coefficient particle vector is $\{h_0, h_1 \dots h_N\}$. The particle vectors are distributed in a D - dimensional search space, where $D = N+1$ for the case of N th order FIR filter. The frequency response of the FIR digital filter can be calculated as,

$$H(e^{jw_k}) = \sum_{n=0}^N h(n)e^{-jw_k n}, \quad (2)$$

where $H(e^{jw_k})$ is the Fourier transform complex vector. This is the FIR filter frequency response. The frequency is sampled in $[0, \pi]$ with M sampling points; the position of each particle vector in D -dimensional search space represents the same coefficients $h(n)$ of the transfer function (1).

In this paper, the authors have adopted a new fitness function in order to achieve higher stop band attenuation and lower stop band ripple and to have an accurate control on the transition width. The fitness function used in this paper is given in (3). Using (3), it is found that the proposed filter design approach results in considerable improvement over PM and other optimization techniques.

$$J_2 = \sum abs[abs(|H(\omega)| - 1) - \delta_p] + \sum [abs(|H(\omega)| - \delta_s)] \tag{3}$$

where *abs* or $|\ |$ indicates the absolute value. For the first term of (3), $\omega \in$ pass band including a portion of the transition band and for the second term of (3), $\omega \in$ stop band including a portion of the transition band. The error function given in (3) represents the generalized fitness function to be minimized using the evolutionary algorithms RGA, conventional particle swarm optimization (PSO), and IPSO individually. Each algorithm tries to minimize this error fitness J_2 and thus improves the filter performance. Unlike other error fitness functions as given in [8] [11] [12] which consider only the maximum errors, J_2 involves summation of all absolute errors for the whole frequency band, and hence minimization of J_2 yields higher stop band attenuation and lesser pass and stop band ripples. Transition width is affected a little. Since the coefficients of the linear phase filter are matched, the dimension of the problem is thus reduced by a factor of 2. By only determining half of the coefficients, the filter can be designed. This greatly reduces the computational burdens of the algorithms, applied to the design of linear phase FIR filters.

3 Evolutionary Techniques Employed

3.1 Real Coded Genetic Algorithm (RGA)

Steps of RGA as implemented for optimization of $h(n)$ coefficients are adopted from [13]. In this work, initialization of real chromosome string vectors of n_p population, each consisting of a set of $h(n)$ coefficients is made. Size of the set depends on the number of coefficients in a particular filter design.

3.2 Particle Swarm Optimization (PSO)

PSO is a flexible, robust population-based stochastic search/optimization technique with implicit parallelism, which can easily handle with non-differential objective functions, unlike traditional optimization methods. Mathematically, velocities of the particles are modified according to the following equation:

$$V_i^{(k+1)} = w * V_i^k + C_1 * rand_1 * (pbest_i^k - S_i^k) + C_2 * rand_2 * (gbest^k - S_i^k) \tag{4}$$

where V_i^k is the velocity of i^{th} particle at k^{th} iteration; w is the weighting function; C_1 and C_2 are the positive weighting factors; $rand_1$ and $rand_2$ are the random numbers between 0 and 1; S_i^k is the current position of i^{th} particle at k^{th} iteration;

$pbest_i^k$ is the personal best of the i^{th} particle at the k^{th} iteration; $gbest^k$ is the group best of the group at the k^{th} iteration. The searching point in the solution space may be modified by the following equation:

$$S_i^{(k+1)} = S_i^k + V_i^{(k+1)} \quad (5)$$

The first term of (4) is the previous velocity of the particle. The second and third terms are used to change the velocity of the particle. Without the second and third terms, the particle will keep on “flying” in the same direction until it hits the boundary. Namely, it corresponds to a kind of inertia represented by the inertia constant, w and tries to explore new areas.

3.3 Improved Particle Swarm Optimization (IPSO)

The global search ability of traditional PSO is very much enhanced with the help of the following modifications. This modified PSO is termed as IPSO [14].

i) The two random parameters $rand_1$ and $rand_2$ of (4) are independent. If both are large, both the personal and social experiences are over used and the particle is driven too far away from the local optimum. If both are small, both the personal and social experiences are not used fully and the convergence speed of the technique is reduced. So, instead of taking independent $rand_1$ and $rand_2$, one single random number r_1 is chosen so that when r_1 is large, $(1 - r_1)$ is small and vice versa. Moreover, to control the balance of global and local searches, another random parameter r_2 is introduced. For birds flocking for food, there could be some rare cases that after the position of the particle is changed according to (4), a bird may not, due to inertia, fly toward a region at which it thinks is most promising for food. Instead, it may be leading toward a region which is in the opposite direction of what it should fly in order to reach the expected promising regions. So, in the step that follows, the direction of the bird’s velocity should be reversed in order for it to fly back into promising region. $sign(r_3)$ is introduced for this purpose. Both cognitive and social parts are modified accordingly. Other modifications are described below.

ii) A new variation in the velocity expression (4) is made by splitting the cognitive component (second part of (4)) into two different components. The first component is called *good experience component*. That is, the particle has a memory about its previously visited best position. This component is exactly the same as the cognitive component of the conventional PSO. The second component is given the name *bad experience component*. The bad experience component helps the particle to remember its previously visited worst position. The inclusion of the worst experience component in the behavior of the particle gives additional exploration capacity to the swarm. By using the bad experience component, the bird (particle) can bypass its previous worst position and always try to occupy a better position.

Finally, with all modifications, the modified velocity of the i^{th} particle vector at the $(k+1)^{\text{th}}$ iteration is expressed as (6).

$$\begin{aligned}
 V_i^{(k+1)} = & r_2 * \text{sign}(r_3) * V_i^k + (1 - r_2) * C_1 * r_1 * \{pbest_i^k - S_i^k\} + \\
 & (1 - r_2) * C_2 * (1 - r_1) * \{gbest^k - S_i^k\} + (1 - r_2) * c_1 * r_1 * (S_i^k - pworst_i^k)
 \end{aligned}
 \tag{6}$$

where $\text{sign}(r_3)$ is a function defined as:

$$\begin{aligned}
 \text{sign}(r_3) = & -1 \quad \text{where } r_3 \leq 0.05 \\
 & = 1 \quad \text{where } r_3 > 0.05
 \end{aligned}$$

V_i^k is the velocity of the i^{th} particle at the k^{th} iteration; r_1 , r_2 and r_3 are the random numbers between 0 and 1; S_i^k is the current position of the i^{th} particle at the k^{th} iteration; $pbest_i^k$ and $pworst_i^k$ are the personal best and the personal worst of the i^{th} particle, respectively ; $gbest^k$ is the group best among all pbests for the group. The searching point in the solution space is modified by (5) as usual.

4 Results and Discussions

4.1 Analysis of Magnitude Response of Band Pass FIR Filter

In order to demonstrate the effectiveness of the proposed filter design method, FIR filters are constructed using PM, RGA, PSO, IPSO algorithms. The MATLAB simulation has been performed extensively to realize the band pass FIR filter of the order of 20. Hence, the length of the filter coefficient is 21.

Table 1. RGA, PSO, IPSO parameters

Parameters	RGA	PSO	IPSO
Population size	120	25	25
Iteration Cycle	600	350	200
Crossover rate	1	-	-
Crossover	Two Point Crossover	-	-
Mutation rate	0.01	-	-
Mutation	Gaussian Mutation	-	-
Selection	Roulette	-	-
Selection Probability	1/3	-	-
C_1	-	2.05	2.05
C_2	-	2.05	2.05
v_i^{\min}	-	0.01	0.01
v_i^{\max}	-	1.0	1.0
w_{\max}	-	1.0	-
w_{\min}	-	0.4	-

The sampling frequency has been chosen as $f_s = 1\text{Hz}$. Also, for all the simulations the number of sampling points is taken as 128. Algorithms are run for 50 times to get the best solutions. The best results are reported in this work. Table 1 shows the best chosen parameters used for different heuristic optimizations algorithms. The algorithm

parameters have been so chosen by many trials to get the best possible results. The parameters of the filters to be designed are: pass band ripple (δ_p) = 0.1, stop band ripple (δ_s) = 0.01. For band pass filter, lower stop band (normalized) edge frequency (ω_{sl}) = 0.25, lower pass band (normalized) edge frequency (ω_{pl}) = 0.35; higher pass band (normalized) edge frequency (ω_{ph}) = 0.65; higher stop band (normalized) edge frequency (ω_{sh}) = 0.75; transition width=0.1. Figure 1 shows the magnitude plot for the FIR band pass (BP) filter of the order of 20. The best optimized coefficients for the designed filters with the order of 20 have been calculated by RGA, PSO and IPSO and given in Table 2.

Table 2. Optimized Coefficients of FIR Band pass Filter of Order 20

h(N)	RGA	PSO	IPSO
h(1)=h(21)	0.028502857888104	0.025165091322598	0.027902253292513
h(2)=h(20)	-0.001893868108392	-0.002300832301629	0.003939814376923
h(3)=h(19)	-0.076189026154460	-0.072788608489910	-0.075465671961181
h(4)=h(18)	0.000994123920259	0.001962570281435	-0.005029715870387
h(5)=h(17)	0.053196793860741	0.055176711318501	0.055687299039737
h(6)=h(16)	-0.000639149080848	0.001324681739036	-0.000020082854816
h(7)=h(15)	0.100057194730152	0.095808566461517	0.096334206843426
h(8)=h(14)	0.001409980793664	-0.003567579531894	0.004760681054475
h(9)=h(13)	-0.299380312728113	-0.297783056747784	-0.298190578247113
h(10)=h(12)	-0.000752480372393	0.001915220607266	-0.002697869505629
h(11)	0.400369877077545	0.400369877077545	0.400369877077545

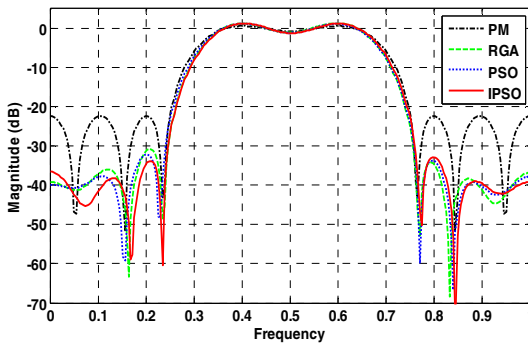


Fig. 1. Magnitude (dB) Plot of the FIR Band pass Filter of Order 20

Table 3 shows the maximum stop band attenuation (dB), maximum pass band ripple (normalized), maximum stop band ripple (normalized) and transition width for all the aforementioned optimization algorithms. From the figures and tables, it is evident the proposed filter design approach IPSO produces higher stop band attenuation and smaller stop band ripple compared to that of PM, RGA and PSO. The filter designed by the IPSO algorithm has a similar transition band response to that of the response produced by RGA and PSO algorithms. For the stop band region, the filters designed by the IPSO method results in the improved responses than the other.

4.2 Comparative Effectiveness and Convergence Profiles of RGA and IPSO

In order to compare the algorithms in terms of the convergence speed, Figure 2 shows the plot of minimum error values against the number of iteration cycles when RGA is employed. Figure 3 shows the plot of minimum error values against the number of iteration cycles when IPSO is employed.

The convergence profiles have been shown for the filter order of 20. From the figures drawn for this filter, it is seen that the IPSO algorithm is significantly faster than the RGA algorithm for finding the optimum filter. The IPSO converges to a much lower fitness in lesser number of iterations. Further, RGA yields suboptimal higher values of error but IPSO yields near optimal (least) error values.

Table 3. Summary of IPSO results with other algorithms for Band pass filter of Order 20

Algorithm	BP filter				
	Maximum stop band attenuation (dB)	Maximum pass Band ripple (normalized)	Maximum stop band ripple (normalized)	Transition width	Execution Time per 100 cycles
PM	22.37	0.076	0.07602	0.0875	-
RGA	30.75	0.167	0.02885	0.0985	3.6867
PSO	32.1	0.148	0.02485	0.0993	2.6082
IPSO	32.78	0.157	0.02041	0.0991	3.0298

With a view to the above fact, it may finally be inferred that the performance of IPSO technique is better as compared to RGA and PSO in designing the optimal FIR filter. All optimization programs are run in MATLAB 7.5 version on core (TM) 2 duo processor, 3.00 GHz with 2 GB RAM.

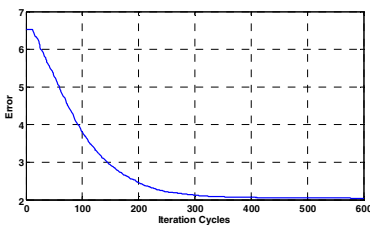


Fig. 2. Convergence Profile for RGA in case of 20th Order Band Pass FIR Filter

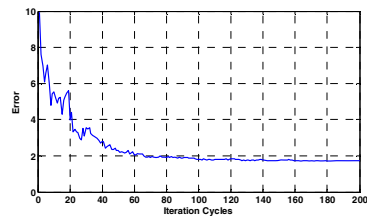


Fig. 3. Convergence Profile for IPSO in case of 20th Order Band pass FIR Filters

5 Conclusions

This paper presents a novel and accurate method for designing linear phase digital band pass FIR filters by using nonlinear stochastic global optimization based on

IPSO. Filters of order 20 have been realized using RGA, PSO as well as the proposed IPSO algorithm. Extensive simulation results justify that the proposed IPSO algorithm outperforms RGA and conventional PSO in the accuracy of the magnitude response of the filter as well as in the convergence speed and is adequate for use in other related design problems.

References

1. Mastorakis, N.E., Gonos, I.F., Swamy, M.N.S.: Design of Two Dimensional Recursive Filters Using Genetic Algorithms. *IEEE Transaction on Circuits and Systems I - Fundamental Theory and Applications* 50, 634–639 (2003)
2. Ahmad, S.U., Antoniou, A.: A genetic algorithm approach for fractional delay FIR filters. In: *IEEE International Symposium on Circuits and Systems, ISCAS 2006*, pp. 2517–2520 (2006)
3. Lu, H.C., Tzeng, S.-T.: Design of arbitrary FIR log filters by genetic algorithm approach. *Signal Processing* 80, 497–505 (2000)
4. Chen, S.: IIR Model Identification Using Batch-Recursive Adaptive Simulated Annealing Algorithm. In: *Proceedings of 6th Annual Chinese Automation and Computer Science Conference*, pp. 151–155 (2000)
5. Karaboga, D., Horrocks, D.H., Karaboga, N., Kalinli, A.: Designing digital FIR filters using Tabu search algorithm. In: *IEEE International Symposium on Circuits and Systems, ISCAS 1997*, vol. 4, pp. 2236–2239 (1997)
6. Karaboga, N.: A new design method based on artificial bee colony algorithm for digital IIR filters. *Journal of the Franklin Institute* 346(4), 328–348 (2009)
7. Kennedy, J., Eberhart, R.: Particle Swarm Optimization. In: *Proc. IEEE Int. Conf. on Neural Network* (1995)
8. Ababneh, J.I., Bataineh, M.H.: Linear phase FIR filter design using particle swarm optimization and genetic algorithms. *Digital Signal Processing* 18, 657–668 (2008)
9. Biswal, B., Dash, P.K., Panigrahi, B.K.: Power quality disturbance classification using fuzzy C-means algorithm and adaptive particle swarm optimization. *IEEE Trans. Ind. Electron.* 56(1), 212–220 (2009)
10. Parks, T.W., McClellan, J.H.: Chebyshev approximation for non recursive digital filters with linear phase. *IEEE Trans. Circuits Theory CT-19*, 189–194 (1972)
11. Luitel, B., Venayagamoorthy, G.K.: Differential Evolution Particle Swarm Optimization for Digital Filter Design. In: *2008 IEEE Congress on Evolutionary Computation (CEC 2008)*, pp. 3954–3961 (2008)
12. Sarangi, A., Mahapatra, R.K., Panigrahi, S.P.: DEPSO and PSO-QI in digital filter design. *Expert Systems with Applications* 38, 10966–10973 (2011)
13. Mandal, D., Ghoshal, S.P., Bhattacharjee, A.K.: Application of Evolutionary Optimization Techniques for Finding the Optimal set of Concentric Circular Antenna Array. *Expert Systems with Applications* 38, 2942–2950 (2010)
14. Mandal, D., Ghoshal, S.P., Bhattacharjee, A.K.: Swarm Intelligence Based Optimal Design of Concentric Circular Antenna Array. *Journal of Electrical Engineering* 10(3), 30–39 (2010)

Text Categorization with K-Nearest Neighbor Approach

Suneetha Manne¹, Sita Kumari Kotha¹, and S. Sameen Fatima²

¹ Department of IT, VRSEC, Vijayawada

² Department of CSE, Osmania University, Hyderabad

suneethamanne74@gmail.com

Abstract. World Wide Web is the store house of abundant information available in various electronic forms. In the past two decades, the increase in the performance of computers in handling large quantity of text data led researchers to focus on reliable and optimal retrieval of information already exist in the huge resources. Though the existing search engines, answering machines has succeeded in retrieving the data relative to the user query, the relevancy of the text data is not appreciable of the huge set. It is hence binding the range of resultant text data for a given user query with appreciable ranking to each document stand as a major challenge. In this paper, we propose a Query based k-Nearest Neighbor method to access relevant documents for a given query finding the most appropriate boundary to related documents available on web and rank the document on the basis of query rather than customary Content based classification. The experimental results will elucidate the categorization with reference to closeness of the given query to the document.

1 Introduction

As the volume of information is getting increased in the internet day by day there is a need for people to have the tools that find, filter the information and manage the resources. It is highly difficult for the people to maintain the huge data manually and it is very time consuming to extract the information effectively without any indexing and classification techniques [6]. Automatic text categorization is one particular tool to retrieve and make use of the text information efficiently.

Over the past two decades, the automatic management of electronic documents has been a major research field in computer science. Text documents have become the most common type of information repositories especially with the increased popularity of the internet and the World Wide Web. Internet and web documents like web pages, emails, newsgroup messages, internet news feed etc., contain million or even billion of text documents [7]. There are several applications where text categorization (classification) plays an important role like technical, professional, business and web based areas. Also the classification is considered to be an important research field used to identify the data and classify it based on several theoretical approaches. Using automatic text categorization the stories can be categorized based on subject categories, academic papers are often classified by technical domains and

sub-domains, patient reports in health care organizations etc. Automatic text categorization is efficient and cheaper when compared to manual categorization where it needs more number of people to manually label or categorize the data. Several methods can be implemented for categorizing the text that varies in their accuracy and computation efficiency [8].

2 Text Categorization Techniques

A large number of statistical classification and machine learning techniques have been applied to text categorization, including regression models, Bayesian classifiers, and Decision Trees, Nearest Neighbor Classifiers, Neural Networks, and Support Vector Machines [9]. Some of the classifiers are discussed in this section.

2.1 Decision Trees

Decision trees are the most widely used inductive learning methods. Their robustness to noisy data and capability to learn disjunctive expressions seem suitable for document classification. One of the most well known decision tree algorithms is ID3 and its successor C4.5 and C5. It is a top-down method which recursively constructs a decision tree classifier. A Decision Tree (DT) text classifier is a tree in which internal nodes are labeled by terms, branches departing from them are labeled by the weight that the term has in the test document, and leafs are labeled by categories. Such a classifier categorize a text document d by recursively test for the weights that the terms labeling the internal nodes have in vector d_j , until a leaf node is reached; the label of this node is then assigned to d_j [10].

A possible method for learning a DT for category C_j consists in a “divide and conquer” strategy of checking whether all the training examples have the same label. If not, selecting a term t_k , partitioning from the pooled classes of the documents that have the same value for t_k , and placing each such class in the same class C_j , which is then chosen as the label for the leaf. The key step is the choice of the term t_k on which to operate the partition. Generally, a choice is made according to an information gain or entropy criterion. However, such a fully grown tree may be prone to over fitting, as some branches may be too specific to the training data. Most DT learning methods thus include a method for growing the tree and one for pruning it, for removing the overly specific branches.

2.4 K-Nearest Neighbor Classifier

The Nearest Neighbor classifier is used for text classification. The Nearest Neighbor classification is a non-parametric method and it can be shown that for large datasets. In this classifier to decide whether the document d_i belongs to the class C_k , the similarity $\text{Sim}(d_i, d_j)$ or Dissim (d_i, d_j) to all documents d_j in the set is determined.

The k most similar training documents are selected. The proportion of neighbors having the same class may be taken as an estimator for the probability of that class and the class with the largest proportions assigned to the document d_j . The algorithm has two parameters (k and similarity/dissimilar value) which decide the performance of the classifier and are empirically determined. However, the optimal number 'k' of neighbors may be estimated from additional training data by cross validation [11] [12]. After having a quick overview of each classifier, it is need to overcome the problems faced by each classification algorithm and develop a new approach to classify the document into a reserved class.

3 Proposed Method

Let p be a document which is represented by a scalar point within a space s . Let s_1, s_2, \dots, s_n be n numbered vector points each representing a training document of the dataset considered as corpora in the same space s . A customary K Nearest Neighbor method requires an input integer k which is the number of vectors retrieved that are relevant to the given point p . K -NN also need a metrics to measure the closeness of each training vector point to the test point p . The training datasets are themselves pre-labeled into predefined category. The result is the label name of maximum number of documents which are close to p of the k points. But when discussing about Information Access Systems (IAS) like Web Search Engines, Answering Machines, Blogs, Software resource providers, etc. the testing input is not in the form of document but a small query in terms of word or phrase itself. The result would be more or less acceptable when dealt with an accurate word. But when the query is in the form of phrase, it may lead to trouble the IAS. We will then segment the given l worded query into l words. For each word in the query, we will calculate the closeness to each training documents and hence by consider k -documents for each word. As a next step, we introduce a new technique— stroking where we build possible phrases among the l -words. Again K -Nearest Neighbor Algorithm is applied on the l batched k documents to further formalize the $l \times k$ documents into phrased numbered of documents. Such iteration of KNN in multi-steps is applied until a single phrase query forms k documents relevant closely to the document.

Ranking plays an important role, as long as search and other information retrieval applications keep developing and growing. Here ranking to each document is a bit easy task as it can be assigned on the basis of basic metric— closeness of the resultant k documents. KNN is considered as the most tractable computationally among most of the Instance Based Classification Methods as it effectively works with huge amount of datasets.

3.1 Text Categorization Overview

Text categorization is the task of automatically assigning input text to a set of categories [1]. The objective of text categorization is to assign a category to an entry from a set of predefined categories to a document. So far many methods of the text

categorization are presented, such as Support Vector Machine, k nearest neighbors, neural network, bayes classier and decision tree, etc. most of them are instance based and some content based. K nearest neighbor is a simple, valid, non-parametric method among them. KNN has undergone into many changes in the present era as the traditional KNN has two fatal defects that are time of similarity computing is huge and its performance depends on training sample set. For multi document text categorization, similarity between unknown samples and also between known samples need to be calculated resulting in the high competency value. Also the test vector matrix becomes high dimensional leading to increase in time complexity. When dealing with query based document categorization, where a single document may serve for multiple categories, a mere binary assignment is not sufficient. For this computation of similarity must be carried out as a special case. As studied in [3], there are three approaches using which we can increase the speed of calculation for KNN:

- By reducing the dimension of text vector,
- By using the smaller sample set,
- By quickening the speed of finding the k nearest neighbors

3.2 Ranking

For a given query, Ranking is one parameter which defines how good a document is better closer than the other document to fall itself under a specific category. Unfortunately the index terms identification which is considered as the most crucial part of ranking will in no way help if considered a Boolean Model. But the Statistical Model is based on similarity between the statistical properties of the text document and the query is no doubt a good and classical approach. In this context too ranking is done predominantly on the basis of three approaches -point wise approach involving classification on the basis of single documents, pair wise approach involving classification of document pairs and list wise approach involving document lists [2]. But, here the proposed KNN approach directly apply statistical model ranking to the text data felicitating the document access time and ranking time. In the next section, the basic structure of the proposed query based text categorization model is discussed.

4 Query Based Classification of Data

In the statistically based vector-space model, a document is conceptually represented by a vector of keywords extracted from the document, with associated weights representing the importance of the keywords in the document and within the whole document collection; likewise, a query is modeled as a list of keywords with associated weights representing the importance of the keywords in the query [4].

Here weighted indexes serve as large number of word attributes enabling for a high precise classification. Once weights are derived from predefined dataset, the input for the query based model is ready to feed. Queries are usually generated from the user's perspective and the more the information available the more is the documents

retrieved. In general, any query consists of a word or a collection of words. Web search engines retrieve the documents related to each word in comparison with the content and show the results on the interface i.e. the browser as a whole. Hence, documents are usually more in number. In Query Based classification system using KNN, the expected minimum inputs are: A query and a positive integral value to how many number of results to be retrieved. The structure is as discussed in the Figure 1.

In Figure 1, S is a vector space consisting of pre-classified four dataset groups labeled L_1, L_2, L_3 and L_4 . each label L_i ($i=1,2,3,4$) contains a finite vector point each representing the document of their respective class. First, for each document in the labeled class, indexes are to be created by classical formalization techniques. Then the given user query is checked for number of words contained in it. Let them be l . This can make the relevancy of document retrieval in an appreciable form. In Figure 1, the user given query is decomposed into l word number of fragments. Then for each fragment of query, the distance between the indexes of each trained vector and query fragment is calculated.

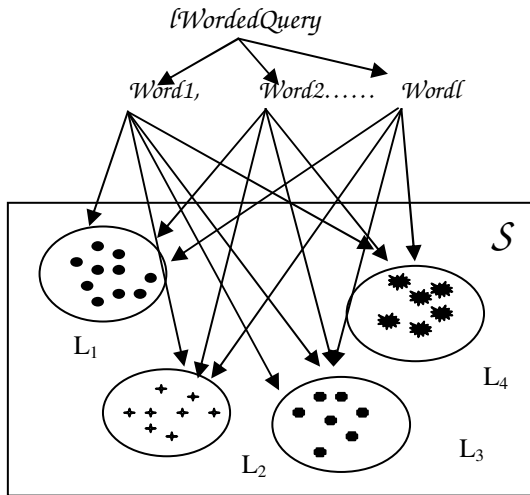


Fig. 1. Query comparing between each of class

There are lots of methods available to calculate the distance. Two important methods are:

1. Euclidean distance: The Root square sum of differences between the two points is considered as distance [6].

$$d_E(l, q) = \sqrt{(l_1 - q_1)^2 + (l_2 - q_2)^2}$$

For a Cartesian coordinate system (two dimensional),

$l = (l_1, l_2)$ is the position of label vector and $q = (q_1, q_2)$ is the position of query word.

2. Manhattan distance: It is the sum of the lengths of the projections of the line segment between the points l, q onto the coordinate axes. It is designed as a n -dimensional vector space [5].

$$d(l, q) = \sum_{i=1}^n |l_i - q_i|$$

Using any of the two methods, we can evaluate the distance between the sub word of query and indexes of each document. By constructing a confusion matrix, it can consider the statistically closer documents. The k valued documents are now obtained for each word in the query. The query terms are then joined into two phrased words and the same above procedure is applied with the earlier retrieved documents in labels as dataset group. The algorithm for such a query based classification is shown in the table 1. The documents in each label are measured with their closeness to the document and least k distances are evaluated and considered for next step of Multilevel KNN approach. The n labeled k documents are now iteratively fed to the same algorithm and processed further so as to get a single query word labeled documents set. Usually modeling training data is time consuming but ranking of documents is simple as the metrics can be built on the basis of distance. The document with less closeness with the document is given first importance. If same distance occurs then document with most phrase build is given most importance.

Table 1. Algorithm for Query based Classification

<p>Input: Query q, i words in it, A preset integral value k, N labeled classes as corpus</p> <p>Algorithm:</p> <ol style="list-style-type: none"> i) Split the given query into i words ii) For int $k=1$ do the steps (iii) and (iv). iii) Calculate dist between l_j where $j = 1, 2, \dots, j$ documents in Label i using <i>distance formula</i> iv) Take k least distant values form label l v) Repeat the steps until l labels vi) Build phrase between i and $i+1^{\text{th}}$ word <p>Do repeat (i) to (vi) for all i worded until l phrase i.e. i worded query</p>
--

5 Implementation Using 20Newsgroups

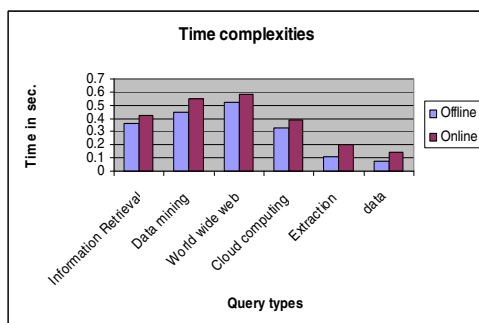
The 20 Newsgroups has been a predominantly using dataset for experiments in text applications of machine learning techniques. The 20Newsgroups is a collection of approximately 20,000 newsgroup documents grouped into 20 predefined categories. The category list is provided in the Table 2. The dataset was originally collected by Ken Lang, for his Newsreader research learning to filter net news paper. The query based KNN approach has different time complexities when implemented for offline and in online. For a given query the processing time at online and offline is observed.

There is comparatively bearable time lag found for the same dataset and is shown in Figure 2. Ranking of the documents retrieved is based on the least value of the distance calculated in the ultimate steps of calculation. KNN application here is done with k value preset to 100.

Table 2. 20Newsgroups dataset

Category	# train docs	# test docs	Total # docs
alt.atheism	480	319	799
comp.graphics	584	389	973
comp.os.ms-windows.misc	572	394	966
Comp.sys.ibm.pc.hardware	590	392	982
Comp.sys.mac.hardware	578	385	963
Comp.windows.x	593	392	985
misc.forsale	585	390	975
rec.autos	594	395	989
rec.motorcycles	598	398	996
rec.sport.baseball	597	397	994
rec.sport.hockey	600	399	999
sci.crypt	595	396	991
sci.electronics	591	393	984
sci.med	594	396	990
sci.space	593	394	987
soc.religion.christian	598	398	996
Talk.politics.guns	545	364	909
Talk.politics.mideast	564	376	940
Talk.politics.misc	465	310	775
Talk.religion.misc	377	251	628
Total	11293	7528	18821

The resultant documents retrieved for the given queries are shown in the Table3. The results have provided a better perspective view for the on KNN approach there by giving a better scope for optimality and feature based categorization too. As the document to be retrieved are predetermined initially (For our experiment k=100), the time complexity may vary accordingly. For huge amount of data, Centroid Based distance calculation gave best results in addition to classical above discussed forms.

**Fig. 2.** Time Complexities chart**Table 3.** Retrieved Documents

Query	Documents retrieved Offline and Online
Information Retrieval	32
Data mining	48
World wide web	52
Cloud computing	15
Extraction	62
Data	70

The performance of the KNN method is tested by selecting different values of parameter k and the number of nearest neighbors. Observed that when $k = m$, KNN becomes comparatively less, where m denotes the number of training queries. Datasets with different k values in terms of different groups as k increases, the performance first increase and then decrease.

6 Conclusions

The method has found the most relevant documents for a given query. Also it is useful for finding the most appropriate boundary to related documents available on web and rank the document on the basis of query rather than customary content based classification. Experimental results shows, this approach of Text Categorization have provided a better perspective view, there by giving a better scope for optimality and feature based categorization. This method has significantly reduced the query response time, improving the accuracy and degree of relevancy.

References

1. Sebastiani, F.: Machine learning in automated text categorization. *ACM Computing Surveys* 34(1), 1–47 (2002)
2. Geng, X., Liu, T.-Y., Qin, T., Arnold, A., Li, H., Shum, H.-Y.: Query Dependent Ranking Using K-Nearest Neighbor. In: *ACM, SIGIR 2008, Singapore, July 20–24 (2008)*
3. Lee, D.L., Chuang, H., Seamons, K.: Document Ranking and the Vector-Space Model, a research thesis (March-April 1997)
4. Guo, G., Wang, H., Bell, D., Bi, Y., Greer, K.: Using kNN Model-based Approach for Automatic Text Categorization
5. Papadopoulos, S., Wang, L., Yang, Y., Papadias, D., Karras, P.: Authenticated Multi-Step Nearest Neighbor Search
6. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Fisher, D.H. (ed.) *Proceedings of ICML-1997, 14th International Conference on Machine Learning*, pp. 412–420. Morgan Kaufmann Publishers, San Francisco (1997)
7. Guru, D.S., Harish, B.S., Manjunath, S.: Clustering of Textual Data: A Brief Survey. In: *The Proceedings of International Conference on Signal and Image Processing*, pp. 409–413 (2009)
8. Al-Shalabi, R., Kanaan, G., Gharaibeh, M.H.: Arabic Text Categorization Using kNN Algorithm
9. Aas, K., Eikvil, L.: Text Categorization: A Survey. Norwegian Computation Center, Oslo (1999)
10. Mitchell, T.M.: *Machine Learning*. McGraw-Hill, Singapore (1997)
11. Hotho, A., Nürnberger, A., Paaß, G.: A Brief Survey of Text Mining. *Journal for Computational Linguistics and Language Technology* 20, 19–62 (2005)
12. Yang, Y., Slattery, S., Ghani, R.: A study of approaches to hypertext categorization. *Journal of Intelligent Information Systems* 18(2), 219–241 (2002)

Extraction Based Automatic Text Summarization System with HMM Tagger

Suneetha Manne¹, Zaheer Parvez Shaik Mohd.¹, and S. Sameen Fatima²

¹ Department of IT, VRSEC, Vijayawada, India

² Department of CSE, Osmania University, Hyderabad

{suneethamanne74, zaheerimpeccable, sameenf}@gmail.com

Abstract. A rough estimation of world's famous search engine Google in year 2010 revealed that the total size of internet has now turned to 2 petabytes. The increase in the performance and fast accessing of web resources has made a new challenge of browsing among huge data on internet. It is hence browsing on web is an under laid topic for researchers. The research on web has turned its steps towards Browsing among Information (BAI) rather than Browsing for Information (BFI). The field of Information Extraction (IE) is offering a huge scope to concise and compact the information enabling the user to decide by mere check at snippets of each link. Automatic text summarization is the process of condensing the source text into a shorter version preserving its information content and overall meaning. In this paper, we propose a frequent term based text summarization technique based on the analysis of Parts of Speech for generating effective and efficient summary.

1 Introduction

Text summarization has been identified as one of research area by the Natural Language Processing community for the past three decades. Summary can be defined as a text that is produced from one or more texts, which conveys important information in the original text(s) with an appreciable compression ratio, significantly less than the source [10]. As the textual data available on the web is increasing apprehensively, the need of automatic text summarization become predominant in the fields of Information Retrieval and Extraction.

Automatic text summarization can be classified into two categories: extractive and abstractive [11]. Extractive summary is a selection of sentences or phrases from the original text with the highest score, without changing the source text. Abstractive summary method uses linguistic methods to examine and interpret the text. Most of the current automated text summarization system use extraction method to produce summary. In this paper, an extractive text summarization system is proposed based on POS tagging by considering Hidden Markov Model using Brown corpus to extract important phrases to build as a summary [12]. Automatic parts of speech tagging, is a well known Machine learning technique that has been addressed by several researchers in the past two decades. Any Natural language has its parts of speech such as verb, noun, adjective...etc. POS tagging (Part of Speech tagging) is a process of

automatic assigning the POS for each word within a given sentence [13]. The two important approaches for POS tagging are Supervised POS Tagging - This type of models require a pre tagged corpus which is used for training to learnt information about the tagger set, word-tag frequencies, rule sets, etc. and Unsupervised POS Tagging - This model does not require any pre tagged corpus. It uses advanced computational techniques to automatically induce tagger sets, transformation rules, etc. Based on this information, they either calculate the probabilistic information needed by the stochastic taggers or induce the contextual rules needed by rule based systems or transformation based systems [9].

2 Excavating from the Past Work

Text Summarization dates back to the year 1958 where H.P. Luhn presented the first exploratory research on automatic methods of obtaining abstracts [1]. According to Luhn, to measure the significance of each sentence within an article, word frequency and sentence scoring are used. A cutoff value for significant factor was initially set depending on which the featured sentences are extracted. But the system restricted itself to only few specific areas of literature arising composition problems and also limited to small input data. In 1989, H.P. Edmunson [2] evaluating the composition problem, proposed a new concept of cue words. He divided the entire structure of the text into two parts. One is —*Body* which contains the main data and second is—*Skeleton* which contains title, heading (e.g. Introduction, Purpose, Conclusions etc.) and format of the file. The Cue words are recognized within the text and are compared with the Cue Dictionary corpus and there by cue weights are calculated. This approach suffered with huge time complexity, lack simplicity.

The first endeavor for generating abstractive summaries was succeeded by ADAM Summarizer [3] in 1975. Rather adopting linguistic techniques, ADAMS is built on the framework of Machine Learning to generate summaries through sentence ranking. It had potential to handle new domains in addition to redundancy elimination. K.R. Mc Keown in his thesis [4] in the year 1984 generated the first summary system using Natural Language Processing (NLP) based on a computational model of discourse strategies of what to say next after a sentence. Rhetorical strategies and a focusing mechanism used in provided a computationally tractable method for generating summaries. In 1995, a research paper [5] presented Term Weighting and Sentence Weighting as important features to recognize the featured sentences. It was succeeded to some extent in solving anaphoric resolutions in summary generated. Barzilay & Elahadad [6], Boguraev & Kennedy [6], in 1997 a common to all these systems is the approach of extracting keyphrases from text as a supervised learning task. All these systems require a separate training document set with keyphrases already assigned in order to function properly. This remained as a challenge for research community. MEAD [7] developed in the year 2001 was a multi document summarization toolkit that implemented multiple summarization algorithms such as position-based, TF×IDF, largest common subsequence, and keywords.

Tokenization: splitting of the sentence into words using StringTokenizer class in the java.util package.

Stop word Removal: Some words are extremely common and occur in a large majority of documents. For example, articles such as “a”, “an”, “the”, “by” appear almost in every text but do not include much semantic information.

Stemming: Stemming refers to identifying the root of a certain word in the document. Any text document, in general contain repetition of same word but with variations in the grammar such as word appearing to be in past, or in present tense and sometimes containing gerund (“ing” suffixed at the end). Stemming is of two types. Derivational Stemming - creating a new word from an existing word, most often by changing the grammatical category. Inflectional Stemming- aims at confining normalized words to regular grammatical variants such as singular or plural or past or present. The stemming rules which are considered in the proposed model are shown in Fig.2.

<i>Penultimate of the Word (n-1)th word</i>	<i>Ends with (suffix)</i>	<i>Replace with</i>
n	ant	“ ”
	ement	
	ment	
	ent	
l	able	
	ible	
i	ic	
a	al	
c	ance	
	ence	
n	ition	
	ication	
e	er	
o	ion	
	ou	
s	ism	

<i>Suffix of the Word</i>	<i>Ends with (suffix)</i>	<i>Replace with</i>
e	icate	“ic”
	ative	“ ”
	alize	“al”
i	iciti	“ic”
l	ical	“ic”
s	ful	“ ”
c	ic	“ ”
n	ition	“y”
	ication	“y”
t	iest	“y”

Fig. 2. Proposed method architecture diagram

The two main advantages of stemming algorithms are space efficiency and retrieval generality. The size of the inverted file can be reduced dramatically. The proposed stemming algorithm is shown in Fig.3.

Lemmatizing: Extracting the commonly featured, same meaning tokenized words so as to avoid repetition (e.g. problems-problem, risks-risk, etc.). The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.


```

Algorithm Stemming (String word)
{
  String StopwordStem(word)
  {
    if word ends with any of ( . : , ; ? ` " ) } ] )
    return word.replace(word.trim(), word.substring(0,
      word.length() - 1));
    else if word start with any of ( { [ ( ` " . , : ; )
    return word.replace (word.trim(),word.substring(1));
  }
  String Stemm(word)
  {
    word = ReplaceStem(word);
    if the word ends with any of( second column of
      table1,2 )
    replace with the respective terms
    return word.replace(word, word.substring(0,
      word.length() - suffix length removed));
  } }

```

Fig. 3. Proposed Stemming Algorithm

3.2 Hidden Markov Model

A Hidden Markov Model is a generalization of a mixture mode where the hidden variables, which control the mixture component to be selected for each observation, are related through a Markov process rather than independent of each other [9]. Both first order (bigram) and second order (trigram) Hidden Markov Model are used for developing the tagger system. Hidden states of the model represent tags and observation states represent words [8]. Equation (1) is used to find the best POS tag sequence in first order HMM (bigram) case.

$$t_{n-1} = \arg \max_{t_1..t_n} P(t_1) \times \prod_{i=2}^n P(t_i | t_{i-1}) \times \prod_{i=2}^n P(w_i | t_i) \tag{1}$$

Equation (2) gives the trigram case

$$t_{n-1} = \operatorname{argmax}_{t_1..t_n} \times P(t_1) \times P(t_2 | t_1) \times \prod_{i=2}^n P(t_i | t_{i-1}, t_{i-2}) \times \prod_{i=2}^n P(w_i | t_i) \tag{2}$$

t_{n-1} is the best POS tag sequence for the input words, $w_1 \cdots w_2$ are sequence of input word, $t_1 \cdots t_2$ are elements of the tag set. $P(t_2 | t_1)$ in equation (2) is also not an ordinary bigram probability. $P(t_1)$ and $P(t_2 | t_1)$ in both equations are conditioned on being the first token in a sentence. In the implementation, we add a dummy POS tag <START> before the first word (two dummies for trigram case). Therefore, $P(t_1)$ in equation (1) is calculated using $P(t_1 | \text{start})$. And also, $P(t_1)$ and

$P(t_2|t_1)$ in equation (2) are calculated respectively using $P(t_1|start, start)$ and $P(t_2|t_1, start)$.

3.3 Chunking and Features in Text Summarization

Extracting high level structures like phrases can be possible using Noun and Verb Chunking. Generally nouns may start with determiners, adjectives, common nouns or pronouns and they continued with any category that may start a noun, or adverbs or punctuation and verbs with verbs, auxiliaries, or adverbs and may be continued with any of the tags, or with punctuation. Rather than running over just the first-best output, we used n-best output. The features in text summarization are as follows.

Term Frequency (TF): The term frequency TF (t, d) of term t in document d is defined as the number of times that t occurs in d. The higher the TF(t,d), the more the term t is representative (or characteristic) of document d.

Term Frequency (TF) = no. of times a term repeated

Inverse Sentence Frequency (ISF): N is the total number of sentences and ni is the number of sentences with term i. Similarly, if the system retrieves terms or phrases then ISF should be replaced with the Inverse Term Frequency (ITF), where N is the vocabulary size, and ni is the number of times a term or phrases appears in the corpus.

$ISF(ti) = \log_2(N) - \log_2(ni) + 1$

Term weight (TW): In order to improve the quality of summaries, we consider term weight (TW) in the process of retrieved documents summarization.

Term Weight (TW) = $[TF * 100] / \text{total numbers of terms in the given document}$

Sentence Length: Sentences that are too short not expected to belong to the summary.

Term Frequency (TF): The tokenized terms after applying various normalized techniques like case folding, stop word removal, stemming, lemmatizing, etc., are now considered as Feature Terms.

Normalized sentence length =

$$\frac{\text{number of words occurring in the sentence}}{\text{number of words occurring in the longest sentence of the document}}$$

Sentence Position: This feature can involve several items, such as the position of a sentence in the document as a whole, it's the position in a section, in a paragraph, etc., and has presented good results in several research projects. If number of sentences in a paragraph be n, then n/2 top sentences are considered top priority than that the next n/2 sentences. Paragraph can be recognized using ends with “//s//s//s//s” (sentence ended with four or more spaces). The final value is normalized to take on values between 0 and 1.

Sentence Weight (SW): Here we compute the weight of each sentence and counts the number of words present in each sentence.

Sentence Weight =

$$\frac{\text{number of featured terms within the sentence}}{\text{total number of terms in a paragraph}} \times 100$$

4 Experimental Results

The proposed system decomposed the given text into its constituent sentences, assigning the POS tag for each word in the text and stores the results in table 1. Unique words are generated and using these words the summary will be presented. Further the feature terms are identified. Finally each sentence is ranked depending on feature terms. The summary will be generated based on weight age of the sentences whose value has 50% and count has 250 for a given sample document. We used 100 texts as training documents and 10 texts as test documents tested against human summary. The Table I shows the word frequency, word weight age and POS forms. Text files from the test set have been selected, and the selected sentences to be in the summary presented in Table 2.

Table 1. Word frequency, count and POS tags

frequency	term	weight	pos	form
1	warehous	12	noun	nns
2	traditiona	18	noun	jj
3	Text	25	noun	nn
4	classifica	31	noun	nn
5	document	37	noun	nn
7	data	50	noun	nn
-----	-----	-----	-----	-----

Table 2. Word frequency, count and POS tags

Paragraph	count	pweight
The process of identifying interesting knowledgeable information from large amounts of databases, data warehouses, or any other information repositories is known as Data Mining.	253	96
Where as Information Retrieval (IR) mainly concerned with the organization and retrieval of Information from a large number of text-based documents.	149	56
Some common information retrieval problems are in general not encountered in traditional database systems, such as unstructured documents, approximate search based on keywords, text mining and the notion of relevance.	218	82
Text categorization labels the web document automatically based on a set of predefined categories. It is observed that people who are involved in research study need to analyze the available research papers, e-books and other resources.	305	115

Best query need to be generated considering the parameters: The sentence with best sentence score, sentences containing words with high TF and TW values, sentence which have optimal sentence position within the paragraph, sentence builds on best POS Tagging criteria and relevance of the (n-1) sentences with the first sentence i.e., the title of the document.

5 Conclusions

In this work, an extractive automatic text summarization approach by sentence extraction using a supervised POS tagging is evaluated. A frequent term based text summarization technique with HMM tagger is designed and implemented successfully in a popular and challenging higher level programming language Java. Ranked sentences are collected by identifying the feature terms and text summary is obtained. Such type of summary generated can also be used for Text Categorization systems in order to label the documents so as to organize easily on web. Thus summarization based on extraction can give a huge scope of study both in Information Extraction systems and Language processing systems.

References

1. Luhn, H.P.: The Automatic Creation of Literature Abstracts. *IBM Journal*, 159–165 (April 1958)
2. Edmundson, H.P.: New Methods in Automatic Extracting. *Journal of the Association for Computing Machinery* 16(2), 264–285 (1969)
3. Pollock, J.J., Zamora, A.: Automatic Abstracting Research at Chemical Abstracts Service. *Journal of Chemical Information and Computer Sciences* 15(4), 226–232 (1975)
4. Brown Tagset,
<http://www.scs.leeds.ac.uk/amalgam/tagsets/brown.html>
5. McKeown, K.R.: *Discourse Strategies for Generating Natural Language Text*. Department of Computer Science, Columbia University, New York (1982)
6. Brandow, R., Mitze, K., Rau, L.F.: Automatic condensation of electronic publications by sentence selection. *Information Processing Management* 31(5), 675–685 (1995)
7. Barzilay, R., Elhadad, M., Boguraev, Kennedy, M.: Using Lexical Chains for Text Summarization. In: *Workshop on Intelligent Scalable Text Summarization*, Ben Gurion University of the Negev, Be'er Sheva (1997)
8. Radev, R., Blair-goldensohn, S., Zhang, Z.: Experiments in Single and Multi-Docuemtn Summarization using MEAD. In: *First Document Understanding Conference*, New Orleans, LA (2001)
9. Karthik Kumar, G., Sudheer, K., Avinesh, P.V.S.: Comparative Study of Various Machine Learning Methods for Telugu Part of Speech Tagging. In: *Proceeding of the NLP AI Machine Learning Competition* (2006)
10. Bahl, L., Mercer, R.L.: Part-Of-Speech assignment by a statistical decision algorithm. In: *IEEE International Symposium on Information Theory*, pp. 88–89 (1976)
11. Gupta, V., Lehal, G.S.: A Survey of Text Summarization Extractive Techniques. *Journal of Emerging Technologies In Web Intelligence* 2(3) (August 2010)
12. Radev, D.R., Hovy, E., McKeown, K.: Introduction to the special issue on summarization. *Computational Linguistics* 28(4), 399–408 (2002)
13. Nahm, U.Y., Mooney, R.J.: Text mining with information extraction. In: *AAAI 2002, Spring Symposium on Mining Answers from Texts and Knowledge Bases* (2002)
14. Nou, C.: *Khmer Part-of-Speech Tagging*. Global Information and Telecommunication Studies. Waseda University
15. Suneetha, M., Sameen Fatima, S.: Corpus based Automatic Text Summarization System with HMM Tagger. *International Journal of Soft Computing and Engineering (IJSCE)* 1(3), 118–123 (2011) ISSN: 2231-2307

Recursive Chain Coding Method for Lossless Digital Image Compression

T. Meyyappan¹, S.M. Thamarai¹, and N.M. Jeya Nachiaban²

¹ Department of Computer Science and Engineering, Alagappa University,
Karaikudi – 630 003, India
{meyslotus, lotusmeys}@yahoo.com

² Department of Computer Science and Engineering,
Thiagarajar College of Engineering, Madurai-9, India
nmjeyan2009@tce.edu

Abstract. Image compression addresses the problem of reducing the amount of data required to represent a digital image. Digital images require large amounts of memory to store and, when retrieved from the internet, can take a considerable amount of time to download. In this paper, the authors propose a new approach to image compression using chain coding which traces contours present in the image. The novelty and better compression ratio of the method is due to its recursiveness in tracing the contours. The proposed method starts with the original image and develop chain codes in a recursive manner, marking the pixels visited earlier and expanding the entropy in eight directions. The proposed method is experimented with sample bitmap images and results are compared with Recursive Crack Coding. The method is implemented in uni-processor machine using C language source code.

Keywords: Contour, Chain Coding, Entropy, Lossless Compression, Lossy compression.

1 Introduction

Compressing an image is significantly different than compressing raw binary data. Of course, general purpose compression programs can be used to compress images, but the result is less than optimal. This is because images have certain statistical properties which can be exploited by encoders specifically designed for them. Also, some of the finer details in the image can be sacrificed for the sake of saving a little more bandwidth or storage space. Lossless compression[2] involves with compressing data which, when decompressed, will be an exact replica of the original data. This is the case when binary data such as executables, documents etc. are compressed[14]. They need to be exactly reproduced when decompressed.

The files that comprise images can be quite large and can quickly take up precious memory space on the computer's hard drive[11]. The size of images can also make downloading from the internet a lengthy process. Transmission of images in their original form increases the time spent in network and we need to increase the bandwidth for fast transmission[3]. On the other hand, compressed images which can be restored at the receiving end can very much reduce network overheads.

The Recursive Chain coding method starts with the original image and develop chain codes in a recursive manner, marking the pixels visited earlier and expanding the entropy in eight directions. The proposed method is applied on various digital images. The performance of the method is compared with 4-direction Recursive Crack Coding. The results are tabulated and plotted.

2 Existing Methods

The four different approaches[4],[6][10] to compression are Statistical Compression, Spatial compression, Quantizing compression, Fractal compression. In spatial approach, image coding is based on the spatial relationship between pixels of predictably similar types. The method proposed in this paper employs spatial approach.

Run-length encoding (RLE) is a very simple form of data compression in which runs of data are stored as a single data value and count, rather than as the original run. This is most useful on data that contains many such run. It is not useful with files that don't have many runs as it could greatly increase the file size. Huffman coding removes coding redundancy. Huffman's procedure creates the optimal code for a set of symbols and probabilities subject to the constraint that the symbols be coded one at a time. When large number of symbols is to be coded, the construction of the optimal binary Huffman code is a difficult task. Arithmetic coding[5] is a form of variable-length entropy encoding used in lossless data compression. Arithmetic coding encodes the entire message into a single number. In predictive coding, information already sent or available is used to predict future values, and the difference is coded. Transform coding[12], on the other hand, first transforms the image from its spatial domain representation to a different type of representation and provides greater data compression compared to predictive methods, although at the expense of greater computational requirements.

3 Connectivity

In many circumstances it is important to know whether two pixels are connected to each other, and there are two major rules[8] for deciding this. Consider a pixel called P, at row i and column j of an image; looking at a small region centered about this pixel, we can label the neighboring pixels with integers[3]. Connectivity is illustrated below:

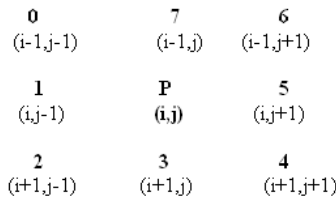


Fig. 1. 8 -Connected Pixels

Two pixels are 8-adjacent if they are horizontal, vertical and diagonal neighbors. The 8-adjacent pixels are said to be connected if they have the same pixel value. Contour coding[5] has the effect of reducing the areas of pixels of the same grey levels to a set of contours that bound those areas. If the areas of same grey level are large with a simple edge, then the compression rate can be very good. In practice, it is best to make all contours circular[5], so that they return to the originating pixel - if necessary along the path that they have already traversed - and to identify the grey level that they lie on and enclose. 8-connected contour[13] is known as chain coding and 4-connected contour is known as crack coding. In this paper, authors used chain coding and grey level of each contour and direction values of the contour alone are stored instead of actual pixel values which take up much storage space.

4 Proposed Method

The proposed method works with the original image as it is. It does not process the image in any way and transform the pixels of the image as in edge detection. It finds all the possible 8-connected contours and stores the 8-directions of the contour along with grey value being examined. The process is repeated with the help of a recursive procedure and marking all the pixels visited along the contour path. The marked pixels are eliminated for further examination of connected pixels. The eight direction chain code values (0 to 7, consuming 3 bits per number) are packed into a byte and stored along with the grey value in output file. No loss of pixels[1] are observed in the proposed compression method. The following is the format of stored compressed image:

Row, Column, Grey-Value, 8-direction chain codes

4.1 Algorithm for Compressing Original Image

The following algorithm shows the sequence of steps to be followed to compress the original image.

- Step 1** Read an uncompressed image file[7]
- Step 2** Read number of rows n and columns m of the image from header
- Step 3** Separate pixels $P[n,m]$
- Step 4** For $i=1$ to n do 5
- Step 5** For $j=1$ to m do
 - Store $P[i,j]$ and its grey value g as beginning of the contour
 - Mark the pixel $P[i,j]$
 - Chain_Code(P,i,j,g)**
- Step 6** Write the header information and contour codes in another file.

Procedure Chain_Code(P,i,j,g)**Begin**

```

if (P[i-1, j-1] equal g) then store 0; Chain_Code(P,i-1, j-1,g);
else if(P[i, j-1] equal g) then store 1; Chain_Code(P,i, j-1,g);
else if(P[i+1, j-1] equal g) then store 2; Chain_Code(P,i+1, j-1,g);
else if(P[i+1, j] equal g) then store 3; Chain_Code(P,i+1, j, g);
else if(P[i+1, j+1] equal g) then store 4; Chain_Code(P,i+1, j+1,g);
else if(P[i,j+1] equals g) then store 5; Chain_Code(P,i, j+1,g);
else if(P[i-1, j+1] equal g) then store 6; Chain_Code(P,i-1, j+1,g);
else if(P[i-1, j] equals g) then store 7; Chain_Code(P,i-1, j,g);
else return;

```

End;**4.2 Algorithm for Restoration of Original Image from Compressed Image**

The following algorithm shows the sequence of steps to restore the original image from compressed image.

- Step 1** Open the compressed image file.
- Step 2** Read number of rows m and columns n of the image from header.
- Step 3** Initialize $P[n,m]$
- Step 4** Repeat steps 5 to 8 until all the chain coded contours are processed
- Step 5** Read starting coordinate position(i, j) and grey value g of next contour.
- Step 6** $P[i, j]=g$;
- Step 7** Read next chain code c ;
- Step 8** Replace_Pixel(P,i, j,g,c);
- Step 9** Write the header information and pixels $P[n,m]$ in another file.

Procedure Replace_Pixel(P,i, j,g,c)**Begin**

```

if(c equals 0) then store P[i-1, j-1]=g;
else if(c equals 1) then store P[i, j-1]=g;
else if(c equals 2) then store P[i+1, j-1]=g;
else if(c equals 3) then store P[i+1, j]=g;
else if(c equals 4) then store P[i+1, j+1]=g;
else if(c equals 5) then store P[i, j+1]=g;
else if(c equals 6) then store P[i-1, j+1]=g;
else if(c equals 7) then store P[i-1, j]=g;
else return;

```

End;

5 Results and Discussion

The authors have developed a package using C language code for the proposed compression and decompression methods. A set of sample bitmap images (both monochrome and color) are tested with the proposed method. The compression percentage varies from 27% to 40% for the samples. No loss of pixels are found while restoring the original image. Original size and compressed size of the images and computation time are plotted. A sample content of the file which stores Starting Position of a pixel, Grey value and Chain Codes of its contour is shown below. 16 bits are required for starting position. 3 bits are needed for storing contour direction. Following numbers represent the direction of contour (one of 8 directions). The last value -1 marks the end of the contour. The last value 999 signifies the end of the file. Every number is packed together into a 3 bits and stored in another file.



Fig. 2(a). Original Bitmap Image

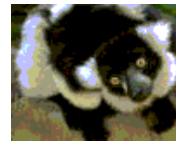


Fig. 2(b). Image after Decompression

Table 1. Experimental Results for Chain Coding and Crack coding

Image No.	Original Size in Bytes	8-Direction Chain Coding				4-Direction Crack Coding			
		Compressed Size in Bytes	Compression %	Compression Time in Seconds	No. of Varying Bytes after Decompression	Compressed Size in Bytes	Compression %	Compression Time in Seconds	No. of Varying Bytes after Decompression
1	9108	5483	40	0.5	0	6416	30	0.8	0
2	8036	5427	32	0.4	0	6270	22	0.8	0
3	8415	5673	33	0.4	0	6392	24	1.3	0
4	7698	5608	27	0.3	0	6271	19	1	0

```

0 0 97 4 3 3 3 3 5 5 4 3 3 5 5 5 5 4 7 7 7 7 1 1 1 1 1 4 3 5 1 0 2 5 -1
0 1 4 7 5 5 4 3 3 4 5 5 5 5 7 1 1 1 1 6 1 6 1 6 1 1 -1
999
    
```

The statistical results are shown below:

No. of bytes in original image = 100 *8=800 bits(A)

No. of bytes in compressed image = (2*8)+(58*3)=190 bits(B)

$$\text{Compression Percentage} = \frac{(A - B)}{A} \times 100 = 76.25\%$$

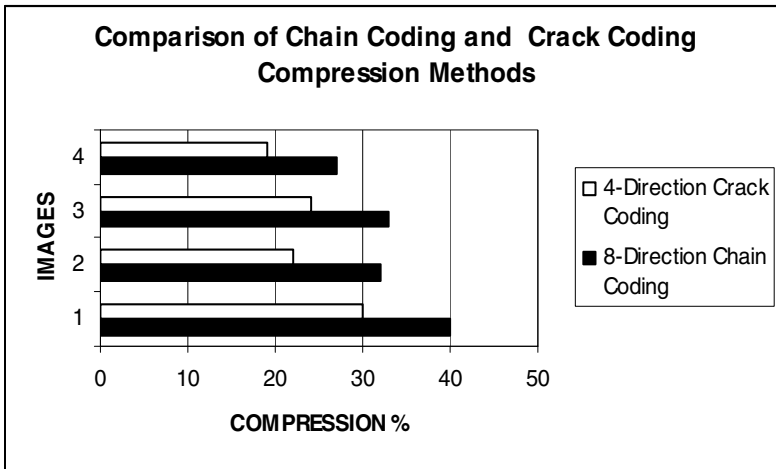


Fig. 3. Performance of the proposed compression method

6 Conclusion

Authors have developed a package using C language and implemented the proposed method. Images compressed with proposed method are restored without any pixel loss. Program execution time, compression percentage are measured for various images. Experimental data are tabulated and plotted. Computation time taken by chain coding is more than the crack coding because of tracing contours with 8-neighboring pixels. Although the proposed method consumes more computation time, the compression percentage is better than the 4-directions Crack Coding method. Parallel processing methods may be adopted to speed up contour tracing operation. The next phase of the research work is in progress to compress images after edge detection.

References

1. Wu, X., Memon, N.: Context-based, adaptive, lossless image coding. *IEEE Trans. Commun.* 45, 437–444 (1997)
2. Ansari, R., Memon, N., Ceran, E.: Near-lossless image compression techniques. *J. Electron. Imaging* 7(3), 486–494 (1998)
3. Meyyappan, T., Thamarai, S.M., Jeya Nachiaban, N.M.: A new method for lossless image compression using recursive crack coding. In: Nagamalai, D., Renault, E., Dhanuskodi, M. (eds.) *DPPR 2011*. CCIS, vol. 205, pp. 128–135. Springer, Heidelberg (2011), doi:10.1007/978-3-642-24055-3; ISSN: 1865-0929, E-ISSN: 1865-0937
4. Ekstrom, M.P.: *Digital Image Processing Techniques (Computational Techniques)*. Academic Press (1984)
5. Low, A.: *Introductory Computer Vision and Image Processing*. McGraw-Hill Publishing Co. (1991)
6. Held, G., Marshall, T.R.: *Data and Image Compression: Tools and Techniques*. Wiley (1996)
7. Miano, J.: *Compressed Image File Formats: JPEG, PNG, GIF, XBM, BMP*. ACM Press (1997)
8. Sayood: *Introduction to Data Compression*, 2nd edn. Academic Press (2000)
9. Jahne, B.: *Practical Handbook on Image Processing for Scientific and Technical Applications*. CRC Press (2004)
10. Parker, J.R.: *Algorithms for Image Processing and Computer Vision*. Wiley (2010)
11. Ames, G.: *Image Compression* (2001)
12. Gautam, B.: *Image Compression using Discrete Cosine Transform & Discrete Wavelet Transform*. In: Department of Computer Science and Engineering, National Institute of Technology, Rourkela (2010)
13. Nelson, C.: *Contour Encoded Compression and Transmission*, M.Sc thesis, Brigham Young University (2006)
14. <http://www.debugmode.com/imagecmp/>

Short Tandem Repeats in Certain Human Genes Reveal a Positive Correlation towards Evolution

Suresh B. Mudunuri¹, Prudhvi Ravi Raja Reddy Mallidi¹, Sujan Patnana¹,
S. Pallamsetty², and Appa Rao Allam³

¹ Department of Computer Science & Engineering, Aditya Engineering College,
Surampalem, East Godavari District, Andhra Pradesh, India 533 437

suresh.mudunuri@aec.edu.in

² Department of Computer Science & Systems Engineering, Andhra University
College of Engineering, Visakhapatnam, Andhra Pradesh, India 530 003

³ CR Rao Advanced Institute of Mathematics, Statistics & Computer Science,
University of Hyderabad Campus, Hyderabad, Andhra Pradesh, India 500 046

Abstract. Mutational dynamics in disease genes are crucial to investigate as many of the genetic diseases occur due to very minor changes in the DNA sequence. Mutations in genomes, especially in gene regions, are the primary reason for the evolution of new organisms. Certain elements in DNA called Short Tandem Repeats (STRs) are known to play a key role in the generation of mutations leading to evolution. In this study, we performed a computational analysis of few disease genes and the impact of mutations in altering the genomic regions of these genes. We observed that the experimentally proven mutations of these genes are resulting in the expansion and contraction of STRs suggesting their positive correlation towards evolution.

Keywords: Short Tandem Repeats, Microsatellites, Bioinformatics, Mutations, Diabetes, Genes, Evolution.

1 Introduction

Short Tandem Repeats (STRs), also known as Microsatellites or Simple Sequence Repeats, are tandem occurrences of repeat motifs of size 1-6 in a DNA sequence [1]. For example, a DNA sub-sequence AGAGAGAGAG is a STR sequence where a motif or unit AG is repeated 5 times, represented as (AG)₅. These repeats are ubiquitous in nature and are spread through out the genomes of almost all organisms. They are responsible for causing several neuro-degenerative diseases in humans and are widely used in DNA Fingerprinting, Forensics, Paternity Studies, Animal Extinction Studies, Population Studies, etc. Studying the role of these Short Tandem Repeats in various human genes would reveal the genomic differences between the diseased and the normal genes and there by understand the reasons for pathogenicity. As STRs are more prone for mutations when compared to other regions of the genome, researchers find them the interesting elements to study. In this study, we have analyzed STRs in four

important genes namely IPF-1, MLH-1, MSH-2 and MSH-6 in humans which are responsible for causing diabetes and related disorders using bioinformatic tools and databases.

Human Insulin Promoter Factor-1 (IPF-1) gene (NM_000209.2) is known to play an important role in pancreas development and in insulin production. Several diseases are linked with this gene as mutations (changes) in this gene sequence are known to cause Diabetes Mellitus and Pancreatic Agenesis [2]. Insulin Promoter Factor 1 (IPF-1) (other synonyms: IDX1, STF-1 and PDX1) is a homeo-domain containing protein, that plays an important role in the transcription of endocrine pancreas specific genes in adults [3][4][5]. Changes in the gene expression of insulin leads to abnormal beta cell function causing diabetes. The remaining 3 human genes MLH-1 (NM_000249.2), MSH-2 (NM_000251.1) and MSH-6 (NM_000179.2) genes are similar to the DNA repair genes mutL and mutS that are found in most bacteria. Studies revealed that failure of these genes to produce the DNA repair enzymes resulted in causing colorectal cancer [6]. The primary role of these genes is to produce proteins that help in DNA mismatch repair during the replication process. As they are very crucial in DNA repairing mechanism, mutations in these genes might cause serious problems in human body. Reports emphasize that if the replication errors are not repaired, they lead to damaged DNA causing colon cancer [7].

We conducted a simple computational analysis of these 4 genes and studied whether there is any correlation between the mutations in these four genes and their STR regions. Moreover, the birth (formation) and death (termination) of these STRs are greatly influenced by the mutations in the DNA. So, we also studied whether the mutations in these genes have any influence in the evolution of STRs.

2 Material and Methods

STRs have been extracted from the four genes (IPF-1, MLH-1, MSH-2 and MSH-6) using the software tool Imperfect Microsatellite Extractor (IMEx Version 2.0) [8][9]. IMEx is used with the following parameters: p%=10; n= Mono: 5, Di: 3, Tri: 2, Tetra: 2, Penta: 2 and Hexa: 2. The remaining parameters are set to default. Translate tool from ExPASy proteomics server [10] has been used to translate the gene sequence into its corresponding protein region in order to map the STRs and the mutations. The translate tool will remove the unnecessary intron regions in the gene sequence.

Further, to find out whether the mutations fall inside the domain regions (regions of functional importance in a gene) or not, we have used the domain prediction tool SMART (Simple Modular Architecture Research Tool) [11]. The experimentally proven mutations of the IPF-1, MLH-1, MSH-2 and MSH-6 genes leading to phenotypic differences were collected from the Human Gene Mutation Database (HGMD) [12]. We did not include mutations in the intron regions and silent mutations (that do not induce any change in the protein sequence). Only those mutations that produced a disease phenotype are considered.

3 Results

The genes IPF-1, MLH-1, MSH-2 and MSH-6 contained 33, 59, 68 and 106 STRs respectively. We mapped the STR regions with the mutations collected from the HGMD database. Out of the 33 STRs in IPF-1 gene, 4 of them are found to be having the mutations that involved in causing pancreatic agenesis and diabetes. In MLH-1 gene, out of the 59 STRs, 29 STRs are mapped to the experimentally proven mutations. In MSH-2 gene, mutations fall in 20 of the 68 STRs. In MSH-6 gene, 18 STRs out of 106 STRs are found to be mapped to the mutations. Though the numbers do not suggest a very positive correlation between STRs and mutations, few interesting results are found during the analysis.

We observed that certain mutations in all these four genes are involved either in the expansion or in the contraction of the STR regions (STR Polymorphism). For example, the mutation InsCCG240 (Proline Insertion) in IPF-1 gene, known to inhibit insulin production, inserts an extra repeat unit CCG expanding the STRs $(CCG)_4$ to $(CCG)_5$ [13]. The extra CCG results in an extra proline in the aminoacid sequence and there by inhibiting the insulin expression to a significant extent [14]. Similarly, the mutation DelTG175 of MSH-2 gene deletes a repeat unit TG that results in contraction of the sequence TG TG TG to TGTG. Many such instances of mutations are listed in the TABLES 1, 2 and 3. Apart from addition or deletion of repeat units, the STR regions are also found to be having point mutations (single nucleotide mutations) in the form of substitutions, insertions and deletions.

Unlike substitution that may or may not change the aminoacid sequence, a single base deletion or insertion will make drastic changes to the final protein sequence. These mutations cause a frame shift in the gene sequence that results in either premature termination of the protein production or change in the entire aminoacid sequence. For example, Pro63fsdel of IPF-1 gene refers to a point mutation (deletion) of a nucleotide C from the 63rd codon leading to a frame shift, there by inducing a stop codon down the line [13]. This mutation falls in a mono-nucleotide repeat $(C)_6$. The frame-shift results in an incomplete protein that has been proved to be responsible for pancreatic agenesis [15].

Among the mutations that fall in STR regions, some mutations clearly indicate their role in causing polymorphism in the STRs. It is indeed interesting to see that some mutations are found to be giving birth (generation) to new STR regions or in other cases, resulting in the expansion of existing STRs. For example, in IPF-1 gene (refer TABLE 1), the mutation C18R substituted the T with C in TG CG CG to make it to CG CG CG. On the other hand, some mutations are disrupting the existing STRs and degenerating them. For example, the mutation DelT891 of MSH-6 gene (refer TABLE 3) deleted the first T from the sequence CT CT CT to degenerate the $(CT)_3$ to $(CT)_2$. The details of the mutations that resulted in the expansion or contraction of STRs are listed in TABLES 1, 2 and 3.

Table 1. List of mutations (substitutions) and their impact in the expansion or contraction of STRs in IPF-1, MLH-1, MSH-2 and MSH-6 genes

Gene	Mutation	Details	Normal	Mutant	Change
IPF-1	C18R	TGC (Cys) to CGC (Arg)	TGCGCG	CG CG CG	+
MLH-1	G22A	GGG (Gly) to GCG (Ala)	GCGGCGGGG	GCG GCG GCG	+
MLH-1	K196STOP	AAA (Lys) to TAA (Stop)	AAAAAA	AAATAA	-
MLH-1	H264R	CAT (Pro) to CGT (Arg)	CATCGTCGT	CGT CGT CGT	+
MLH-1	P496L	CCC (Pro) to CTC (Leu)	CCCCCC	CCCTCC	-
MLH-1	N551T	AAC (Asn) to ACC (Thr)	AACACCACC	ACC ACC ACC	+
MLH-1	K618T	AAG (Lys) to ACG (Thr)	AAG AAG AAG	AAGACGAAG	-
MSH-2	L93F	CTT (Leu) to TTT (Phe)	TCTTTT	TTTTTT	+
MSH-2	P868A	CCA (Pro) to CGA (Ala)	CCAGCAGCA	GCA GCA GCA	+
MSH-6	C559Y	TGC (Cys) to TAC (Tyr)	TG TG TG C	TGTGTAC	-
MSH-6	P1087T	CCC (Pro) to ACC (Thr)	CCCCCCCC	CCCCCACC	-

Table 2. List of insertions and their impact in the expansion or contraction of STRs in IPF-1, MLH-1, MSH-2 and MSH-6 genes

Gene	Mutation	Normal	Mutant	Change
IPF-1	InsCCG240	CCG CCG CCG CCG CCG CCG	ccg CCG CCG	+
MLH-1	InsAA117	AAAA	AAaaAA	+
MLH-1	InsA166	AAAAA	AAaAAA	+
MLH-1	InsT289	AAAAA	AtAAAA	-
MLH-1	InsC495	CCCCCC	CCCcCCC	+
MLH-1	InsT506	TTTT	TTtTT	+
MLH-1	InsTTATA547	TTATA	TTATA ttata	+
MLH-1	InsCA717	CACA	CA ca CA	+
MSH-2	InsG61	GGGGG	GGgGGG	+
MSH-2	InsGG67	GGGG	GGggGG	+
MSH-2	InsT84	TTTT	TTtTT	+
MSH-2	InsA228	AAAAAAA	AAAAaAAA	+
MSH-2	InsAT299	ATAT	AT at AT	+
MSH-2	InsA422	AAAAA	AAAaAA	+
MSH-2	InsA566	AAAA	AAaAA	+
MSH-2	InsA881	AAAAAA	AAAaAAA	+
MSH-6	InsGTGA653	GTGA	GTGA gtga	+
MSH-6	InsATTA871	ATTA	ATTA atta	+
MSH-6	InsTA1065	TATA	TA ta TA	+
MSH-6	InsC1085	CCCCCCCC	CCCCcCCCC	+
MSH-6	InsT1107	TTTT	TTtTT	+
MSH-6	InsGTCA1329	GTCA	GTCA gtca	+
MSH-6	InsTTGA1356	TTGA	TTGA ttga	+

Table 3. List of deletions and their impact in the expansion or contraction of STRs in IPF-1, MLH-1, MSH-2 and MSH-6 genes

Gene	Mutation	Normal	Mutant	Change
IPF-1	DelC63	CCCCC _c	CCCCC	-
MLH-1	DelG20	CGCGgCG	CG CG CG	+
MLH-1	DelG21	GGgGG	GGGG	-
MLH-1	DelA166	AAAaAA	AAAAA	-
MLH-1	DelGA198	GAgaGA	GAGA	-
MLH-1	DelAA285	AAaaA	AAA	-
MLH-1	DelA447	GGaGGGGG	GGGGGGG	+
MLH-1	DelA495	CCCcCC	CCCCC	-
MLH-1	DelAAG616	AAGaagAAG	AAGAAG	-
MLH-1	DelC647	CCcCC	CCCC	-
MLH-1	DelCACA726	CA ca ca	CA	-
MLH-1	DelCA726	CA ca CA	CACA	-
MSH-2	DelG61	GGgGG	GGGG	-
MSH-2	DelTG175	TGtgTG	TGTG	-
MSH-2	DelA227	AAAaAAA	AAAAAA	-
MSH-2	DelA247	AAaAA	AAAA	-
MSH-2	DelAA564	AAaaA	AAA	-
MSH-2	DelAG876	AGagAG	AGAG	-
MSH-2	DelA881	AAAaAA	AAAAA	-
MSH-6	DelAGAGA378	AGAGA agaga	AGAGA	-
MSH-6	DelAGG383	AGG agg AGG	AGGAGG	-
MSH-6	DelT891	Ct CT CT	CCTCT	-
MSH-6	DelC1085	CCCCcCCC	CCCCCCC	-
MSH-6	DelTT1102	TTTtTT	TTTTT	-
MSH-6	DelT1102	TTTtTTT	TTTTTT	-

The + / - in these tables indicate whether the STR is expanded or contracted. The three tables show that insertions resulted in more number of expansions where as deletions degenerated many of the STR regions.

Apart from point mutations, it is also found that insertions or deletions containing more than one nucleotide occur resulting in the generation or degeneration of STR regions. For example, InsTTATA547 mutation of MLH-1 gene gave birth to a new STR (TTATA)₂. Similarly, DelAGAGA378 of MSH-6 gene resulted in the degeneration of (AGAGA)₂ to AGAGA. These mutational dynamics that result in the expansion and contraction of STRs indicate a positive correlation towards the expansion or contraction of genomes that lead to evolution. Lastly, we mapped these mutations with the domain regions (refer TABLE 4) extracted using SMART domain prediction server. Only 11 mutations occur in the domain regions out of the 55 mutations that caused STR polymorphism. It is interesting to note that though many of these STR mutations fall outside the domain regions, they resulted in diseases. This is in contrast to the observation that mutations in domain region effect the protein function. Our results suggest

Table 4. List of domains in IPF-1, MLH-1, MSH-2 and MSH-6 genes

Gene	Domain	Start	End
IPF-1	Homeodomain	146	208
MLH-1	Hatpase_c	23	158
MSH-2	MUTsd	321	645
MSH-2	MUTsac	662	849
MSH-6	MUTsd	753	1102
MSH-6	MUTsac	1127	1321

that though the mutations occur outside the domain regions, they are known to inhibit the protein production.

4 Conclusion

Our analysis indicate that STRs and mutations have an association among themselves and the results indicate that they play a significant role in the evolution of genomes. We found that certain mutations are responsible for the expansion or contraction of STRs leading to the change in the genome size on a whole. Some of the point mutations in STR regions are found to be causing a frame-shift there by resulting in abnormal protein formation irrespective of whether they fall in a domain region or not. It is observed that some mutations are giving birth to a new STR region by adding new repeat unit where as some deletions are degenerating them. Hence, we conclude that STR regions are made dynamic due to the mutations and over a period of time, these changes might result in the change in the genome sequence on a whole which is the basis for evolution. The study has been performed with only few genes and extending this work on a large scale by analyzing large number of genes might give a more conclusive indication of the role of STRs and mutations in evolution.

Acknowledgments. The authors would like to thank Dr. T.V.S.Girendranath, Dean, CSE & IT and the management of Aditya Engineering College for their support and for providing necessary infrastructure for carrying out this work. The efforts of our under-graduate students, Aparna Sampara and Srikanth Reddy Kovvuri, are greatly acknowledged.

References

1. Tautz, D., Schlotterer, C.: Simple Sequences. *Current Opinion in Genetics & Development* 4, 832–837 (1994)
2. Kim, S.K., Selleri, L., Lee, J.S., Zhang, A.Y., Gu, X., Jacobs, Y., Cleary, M.L.: Pbx 1 inactivation disrupts pancreas development and in *Ip1* deficient mice promotes diabetes mellitus. *Nature Genetics* 30, 430–435 (2002)
3. McKinnon, C.M., Docherty, K.: Pancreatic duodenal homeobox–1, PDX-1, a major regulator of beta cell identity and function. *Diabetologia* 44, 1203–1214 (2001)

4. Petersen, H.V., Serup, P., Leonard, J., Michelsen, B.K., Madsen, O.D.: Transcriptional regulation of the human insulin gene is dependent on the homeodomain protein STF1/IPF1 acting through the CT boxes. *Proceedings of the National Academy of Sciences* 91, 10465–10469 (1994)
5. Ohlsson, H., Karlsson, K., Edlund, T.: IPF1, a homeodomain-containing transactivator of the insulin gene. *The EMBO Journal* 12, 4251–4259 (1993)
6. Lawes, D.A., Pearson, T., Sengupta, S., Boulos, P.B.: The role of MLH1, MSH2 and MSH6 in the development of multiple colorectal cancers. *British Journal of Cancer* 93, 472–477 (2005)
7. Peltomäki, P.: Deficient DNA mismatch repair: a common etiologic factor for colon cancer. *Human Molecular Genetics* 10, 735–740 (2001)
8. Mudunuri, S.B., Nagarajaram, H.A.: IMEx: Imperfect Microsatellite Extractor. *Bioinformatics* 23, 1181–1187 (2007)
9. Mudunuri, S.B., Kumar, P., Rao, A.A., Pallamsetty, S., Nagarajaram, H.A.: G-IMEx: A comprehensive software tool for detection of microsatellites from genome sequences. *Bioinformatics* 5, 001–003 (2010)
10. Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R.D., Bairoch, A.: ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Research* 31, 3784–3788 (2003)
11. Schultz, J., Milpetz, F., Bork, P., Ponting, C.P.: SMART, a simple modular architecture research tool: identification of signaling domains. *Proceedings of the National Academy of Sciences* 95, 5857–5864 (1998)
12. Krawczak, M., Ball, E.V., Fenton, I., Stenson, P.D., Abeyasinghe, S., Thomas, N., Cooper, D.N.: Human Gene Mutation Database: biomedical information and research resource. *Human Mutation* 15, 45–51 (2000)
13. Allam, A.R., Suresh, B.M.: Computational Analysis of Microsatellites in Human Insulin Promoter Factor 1 Gene. *J. Proteomics Bioinformatics* 1, 001–004 (2008)
14. El Habib Hani, D.A.S., Chèvre, J.C., Durand, E., Stanojevic, V., Dina, C., Habener, J.F., Froguel, P.: Defective mutations in the insulin promoter factor-1 (IPF-1) gene in late-onset type 2 diabetes mellitus. *Journal of Clinical Investigation* 104, R41–R48 (1999)
15. Stoffers, D.A., Zinkin, N.T., Stanojevic, V., Clarke, W.L., Habener, J.F.: Pancreatic agenesis attributable to a single nucleotide deletion in the human IPF1 gene coding sequence. *Nature Genetics* 15, 106–110 (1997)

Statistical Approach Based Keyword Extraction Aid Dimensionality Reduction

M. Ramakrishna Murty¹, J.V.R. Murthy²,
P.V.G.D. Prasada Reddy³, and Suresh Chandra Satapathy⁴

¹ Dept of CSE, GMR Institute of Technology,
Rajam, Srikakulam(Dist) A.P., India
ramakrishna.m@gmrit.org

² Dept of CSE, JNTU, Kakinada, A.P., India
mjonnalagedda@gmail.com

³ Dept of CS&SE, A.U, Visakhapatnam, A.P., India
prasadreddy.vizag@gmail.com

⁴ Dept of CSE, ANITS, Visakhapatna, A.P., India
sureshsapathy@gmail.com

Abstract. In the text document analysis process keywords are often represented in bag-of-words or vector space model. This representation is high-dimensional and sparse. Keyword extraction is considered as core technology of all automatic processing for text materials. Keywords represent in condensed form the essential content of a document. In this paper we used keyword extraction techniques for find an index terms that contain most important information and unique identify the documents. We proposed keyword extraction based text summarization techniques helps to reduce dimensionality of the vector space model at initial level.

Keywords: Dimensionality, Entropy, Keywords, Co-occurrence.

1 Introduction

Large collections of documents are becoming increasingly common. Such large document collection is important to organize them into meaningful patterns for further analysis. [2] Standard text mining and information retrieval techniques of text documents usually rely on word matching. Keywords represent in condensed form the essential content of a document. [6] Keywords can also be used to enrich the presentation of search results. Identify keywords to be used in extractive summarization of text documents.

Keywords are the index terms that contain the most important information about the contents of the document. Automatic keyword extraction is the task to identify a small set of words, key phrases or keywords from a document that can describe the meaning of document. Users are decided to whether the document is relevant for them or not based on keywords. [4] Keywords resemble as a brief summaries of the document. Everyday so many numbers of articles, books, papers are published which makes it very hard to go through all these text material; instead there is a need of good

information extraction or summarization methods which provide the actual contents of a given document.

Keyword is the smallest unit which expresses meaning of entire document. This keyword is used for many applications and take advantage of it such as indexing, text summarization, information retrieval, document classification, text clustering, filtering, cataloging, topic detection and tracking, information visualization, report generation, web searches, etc.

Keyword Extraction methods are divided into four categories, namely Statistics approach, Linguistics approach, Machine Learning approaches and other approaches.[7]

1.1 Statistics Approach

Statistics methods are simple may be efficient in computation. Statistics-based approaches derive weights of key terms and determine the sentence importance by the total weight the sentence contains. The statistical information of the words can be used to identify the keywords in the document. [2] Statistical techniques are themselves on term frequency to determine the term importance. Sentences with more important terms are extracted in higher priorities. Common ways to determine term importance include TF-IDF, entropy, mutual information, and statistics. Sometimes, term importance is strong if the terms belong to title words, cue-phrases, and/or capitalized words. Moreover, sentence importance can also be adjusted according to its length and where it locates in the document.

Cohen uses N-Gram statistical information to automatically index the document. N-Gram is language and domain independent. Other statistical methods include word frequency, TF*IDF, word co-occurrence, etc.

1.2 Linguistic Approach

Linguistics-based approaches identify term relationship in the document through part-of-speech tagging, grammar analysis, thesaurus usage, and extract meaningful sentences. In the linguistic approach the linguistic features of the words mainly sentences and documents. [12]The linguistic approach includes the lexical analysis, syntactic analysis discourse analysis and so on. Linguistic approaches look into term semantics, which may yield better summary results.

1.3 Machine Learning Approaches

The third approach is machine learning approach, in which Keyword Extraction can be seen as supervised learning, Machine Learning approach employs the extracted keywords from training documents to learn a model and applies the model to find keywords from new documents. This approach includes Naïve Bayes, Support Vector Machine, etc.

Other approaches about keyword extraction mainly combines the methods mentioned above or use some heuristic knowledge in the task of keyword extraction, such as the position, length, layout feature of words, html tags around of the words, etc[13].

2 Preprocessing

Typically, a large number of words exist in even a moderately sized set of documents where a few thousand words or more are common. Thus for large document collections, both the row and column dimensions of the matrix are quite large. So our work is to identify the mostly weighted words are called as keywords for the document that reduce the size of the matrix.

Before applying techniques to get keywords, preprocess the document. The preprocess method gives help to make our keyword extraction easy and it indirectly helps to reduce the dimensionality of vector space model. In general text documents contain set of sentences each sentences contains words, all words may not carry the meaning. Those meaningless words are called as stopwords. First eliminate *stopwords*, special symbols, and commas in the document.

The stopword elimination incorporates that into a porter algorithm for stemming that gives effective preprocessing of document. [1] The Porter stemmer is divided into five steps, in step1 removes the *i*-suffixes and step 2 to 4the *d*-suffixes. Composite *d*-suffixes are reduced to single *d*-suffixes one at a time. So for example if a word ends *icational*, step 2 reduces it to *icate* and step 3 to *ic*. Three steps are sufficient for this process in English. Step 5 does some tidying up.

The basic idea is to represent each document as a vector of certain keyword word frequencies. In order to do so, the following parsing, stemming steps are needed.

1. Extract all unique words from the entire set of documents, without consider case.
2. Eliminate “stopwords” which have not content such as “a”, “and”, “the”, etc.
3. In the step outline a simple preprocessing scheme, that find the “root” word and eliminating plurals, tenses, prefixes, and suffixes.
4. Count the frequency occurrences of each word for every document.
- 5 Using information-theoretic criteria eliminate non-content-bearing “high-frequency” and “low-frequency” words. The high frequency words carry information.

3 Related Work

According to our earlier work in the text analysis, preprocessing helps somehow to find the words but those words are not to apply directly to create a vector space model, because those words count is more and it leads to high-dimensionality of the vector space model. We proposed to use the keyword extraction techniques in the text analysis process to find the content bearing words from the document.

In our proposed method we get only high content bearing words. This proposal helps in several ways for further text analysis; one important is that reducing the high dimensionality of the vector space model. We used statistical based keyword extraction techniques for multiple documents.

Keyword extraction is the process of extracting a few salient words (or phrases) from a given text and using the words to represent the text [3]. Text summarization is also the process of creating a compressed version of a given document that delivers the main topic of the document. The two tasks are similar in essence because they both aim to extract concise representations for documents. Automatic text summarization and keyword extraction have drawn much attention for a long time

because they both are very important for many text applications, including document retrieval, document clustering, etc. Particularly the above two tasks are help to reduce the dimensionality of the vector space model.

In this paper we used statistical based keyword extraction methods to summarize documents and it aid for the dimensional reduction.

4 Statistical Methods

Statistics methods are simple may be efficient in computation. Statistics-based approaches derive weights of key terms and determine the sentence importance by the total weight the sentence contains. Statistical techniques base themselves on term frequency to determine the term importance. Sentences with more important terms are extracted in higher priorities. Common ways to determine term importance include TF-IDF, entropy, mutual information, and statistics. Sometimes, term importance is reinforced if the terms belong to title words, cue-phrases, and/or capitalized words. Moreover, sentence importance can also be adjusted according to its length and where it locates in the document.

The statistical information of the words can be used to identify the keywords in the document. Statistical techniques are themselves on term frequency to determine the term importance. [8] Sentences with more important terms are extracted in higher priorities. Common ways to determine term importance include TF-IDF, entropy, mutual information, and statistics.

4.1 Term and Term Weights (TF-IDF)

The task of keyword extraction is to identify a set of words, representative for a document. [1] To achieve this we use a simple statistical approach. Thereby, as we intend to exploit the properties of a document and of a repository, we used a simple statistical measure. One of the simple weighting is TF*IDF. The TF part intends to give a higher score to a document that has more occurrences of a term, while the IDF part is to penalize words that are popular in the whole collection.

The keyword extraction is conducted exploiting the TF*IDF weight of the term. It is calculated according to the formula:

$$TF * IDF (term) = \frac{TFi * \log(\frac{N}{ni})}{\sqrt{\sum_{i=1}^n (TFi)^2 (\log(\frac{N}{ni}))^2}} \quad (1)$$

where $TF(term)$ is the frequency of a term in the given document, N is the total number of documents in the collection, $DF(term)$ is number of documents, that contain the term. Which is the term having more TF-IDF comparative certain threshold consider as keyword for document analysis.

4.2 Entropy Based Method

After filtering and stemming there is also further decrease the number of words that should be used indexing or keyword selection procedure can be used. In this case,

only the selected keywords are used to describe the documents. A simple method for keyword selection is to extract keywords based on their entropy. E.g. for each word t in the vocabulary the entropy can be computed:

$$w(t) = 1 + \frac{1}{\log_2|D|} \sum_{d \in D} p(d, t) \log_2 P(d, t) \tag{2}$$

where $P(d, t) = \frac{tf(d, t)}{\sum_{t=1}^n tf(d, t)}$

Equation (2) gives the entropy of the given word. Here the entropy gives a measure how well a word is suited to separate documents by keyword search. For instance, words that occur in many documents will have low entropy. [5] The entropy can be seen as a measure of the importance of a word in the given domain context. As index words a number of words that have high entropy relative to their overall frequency can be chosen, i.e. of words occurring equally often those with the higher entropy can be preferred to place in the vector space model.

4.3 Co-occurrence

We use co-occurrence of words as the primary way of quantifying semantic relations between words. According to the distributional hypothesis semantically similar words occur in similar contexts, i.e. they co-occur with the same other words [9]. Therefore rather than using the immediate co-occurrence of two terms as a measure for their semantic similarity we will compare the co occurrences of the terms with all other terms. We formalize this intuition by defining a so called co-occurrence distribution of each word which is simply the weighted average of the word distributions of all documents in which the word occurs [10]. We then operationalize the “semantic similarity” of two terms by computing similarity measure(s) for their co-occurrence distributions. The co-occurrence distribution of a word can also be compared with the word distribution of a text. This gives us a measure to determine how typical a word is for a text.

We simplify a document to a bag of words. Thus, consider a set of n term occurrences W each being an instance of a term t in $T = \{t_1 \dots t_m\}$, and each occurring in a source document d in a collection $C = \{d_1, \dots d_M\}$. Let $n(d, t)$ be the number of occurrences of term t in d , $n(t) = \sum n(d, t)$ the number of occurrences t , and $N(d) = \sum_t n(d, t)$ the number of term occurrences in d . We define probability distributions

$$\begin{aligned} Q(d) &= N(d)/n && \text{on } C \\ Q(t) &= n(t)/n && \text{on } T \end{aligned}$$

[9] A document consists of sentences. a sentence is considered to be a set of words separated by a stop mark (“.”, “?” or “!”). We also include document titles, section titles, and captions as sentences. Two terms in a sentence are considered to co-occur once. That is, we see each sentence as a “basket,” ignoring term order and grammatical information except when extracting word sequences. We can obtain frequent terms by counting term frequencies.

First, frequent terms are extracted. Co-occurrences of a term and frequent terms are counted. If a term appears frequently with a particular subset of terms, the term is

likely to have important meaning. The degree of bias of the co-occurrence distribution is measured by the χ^2 -measure.

Assuming that term w appears independently from frequent terms G , the distribution of co-occurrence of term w and the frequent terms is similar to the unconditional distribution of occurrence of the frequent terms [11]. Conversely, if term w has a semantic relation with a particular set of terms $g \in G$, co-occurrence of term w and g is greater than expected; the distribution is to be biased.

$$\chi^2(w) = \sum_{g \in G} \frac{(freq(w,g) - nwp_g)^2}{nwp_g} \tag{3}$$

5 Experimental Results

According to our earlier work of text data analysis, preprocessing step produce keywords we depending on those words and construct the vector space model. The vector space model is high dimensional and sparse. Reduce the dimensionality many methods are used like singular value decomposition, principal component analysis and factor analysis. But here our proposal is before apply any dimensionality reduction technique let us do keyword extraction techniques lend a hand to dimensionality reduction techniques. Our proposal here is that keyword extraction is used for multiple applications. Let us say the first advantage is to summarize the text, and helps to reduce the size of the representation model i.e. vector space model.

In this paper we used statistical approaches for keyword extraction namely TF-IDF, Entropy, and co-occurrence. We experiment with these three method for finding keywords and compare the results. The results are showing that keyword extraction methods helping to dimensionality reduction methods. These keywords are also facilitating to analyze the text data analysis. Fig.1 shows the words reduced using TF-IDF method. Fig.2 shows the words reduced using Entropy based approach. In the Fig.3 shows the comparison of word reduction in both methods.

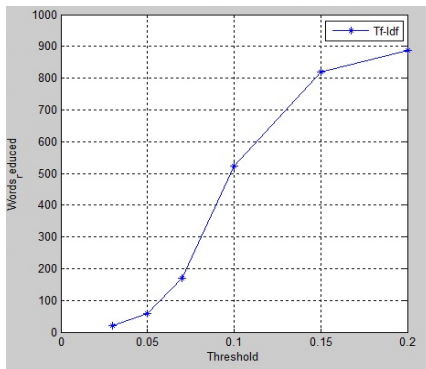


Fig. 1. Reduced words while using TF-IDF for different threshold values

Used standard data set for the experimentation, the dataset has twenty text documents; each method reduced some number of words that is boon for the dimensionality reduction. Entropy based method is best comparatively TF-IDF method. The entropy-based keyword extraction method helps to find the weighed key words from the documents. These weighted keywords used to construct vector space model is optimal and reduced form.

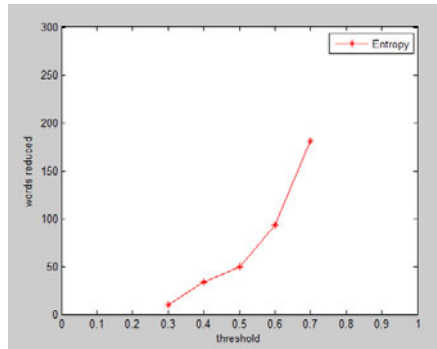


Fig. 2. Reduced words while using Entropy for different threshold levels

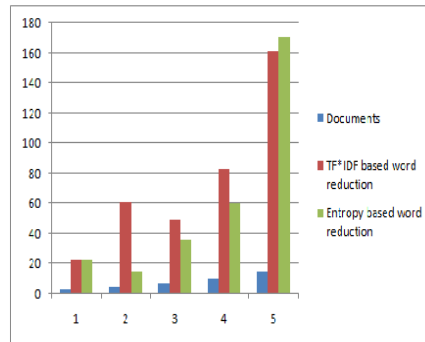


Fig. 3. Comparison of TF*IDF & Entropy word reduction for standard documents

6 Conclusion and Future Work

In this paper we proposed keyword extraction techniques helps to reduce the dimensionality of the vector space model that makes easy to evaluate the text documents. We make comparative study of statistical keyword extraction methods and found that TF-IDF is simple technique, but Entropy method is helps to find the most import words from the document.

Future works include incorporating more features to improve the existing results. The features that can be included are position weight algorithm which weighs words

according to their position of occurrence in the document, length of the word by giving more importance to the lengthy words as compared to the short words, informative feature selection such as bold, italic, underlined words.

References

1. Salton, G., Buckley, C.: Term weighting approaches in automatic text retrieval. Cornell University, Tech. Rep. (1987)
2. Bracewell, D.B., Ren, F.: Multilingual Single Document Keyword Extraction For Information Retrieval. In: Proceedings of NLP-KE, pp. 517–522 (2005)
3. Frank, E., Paynter, G.W., Witten, I.H., Gutwin, C., Nevill-Manning, C.G.: Domain-specific keyphrase extraction. In: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, pp. 668–673 (1999)
4. Azcarraga, A.P., Yap Jr., T.N.: Comparing Keyword Extraction Techniques for WEBSOM Text Archives Singapore 117543 (65) 874-6563 dcsapa@nus.edu.sg
5. Chandra, M., Gupta, V., Pal, S.K.: A Statistical approach for Automatic Text Summarization by Extraction. In: 2011 International Conference on Communication Systems and Network Technologies (2011)
6. Wan, X., Yang, J., Xiao, J.: Towards an Iterative Reinforcement Approach for Simultaneous Document Summarization and Keyword Extraction. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, pp. 552–559. Association for Computational Linguistics (2007)
7. Zha, Y.: Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In: Proceedings of SIGIR 2002, pp. 113–120 (2002)
8. Rafiqul Islam, M., Rakibul Islam, M.: An Improved Keyword Extraction Method Using Graph Based Random Walk Model. In: Proceedings of 11th International Conference on Computer and Information Technology (ICCIT 2008), December 25-27 (2008)
9. Matsuo, Y., Ishizuka, M.: Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information. In: International Journal on Artificial Intelligence Tools World Scientific Publishing Company (2003)
10. Andrade, M., Valencia: Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics* 14(7), 600–607 (1998)
11. Jones, S., Paynter, G.: Automatic extraction of document keyphrases for use in digital libraries: evaluation and applications. *Journal of the American Society for Information Science and Technology* (2002)
12. Matsuo, Y., Ishizuka, M.: Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools* 13(1), 157–169 (2004)
13. Sun, Y.-H., He, P.-L., Chen, Z.-G.: An Improved Term Weighting Scheme For Vector Space Model. In: Proceedings of the Third International Conference on Machine Learning & Cybernetics, Shanghai, August 26-29 (2004)

Hybridization of Rough Set and Differential Evolution Technique for Optimal Features Selection

Suresh Chandra Satapathy¹ and Anima Naik²

¹ Sr Member IEEE, ANITS, Vishakapatnam
sureshsatapathy@ieee.org

² MITS Rayagada
animanaik@gmail.com

Abstract. Dimensionality reduction of a feature set refers to the problem of selecting relevant features which produce the most predictive outcome. In particular, feature selection task is involved in datasets containing huge number of features. Rough set theory has been one of the most successful methods used for feature selection. However, this method is still not able to find optimal subsets. But it can be made to be optimal using different optimization techniques. This paper proposes a new feature selection method based on Rough set theory hybrid with Differential Evolution (DE) try to improve this. We call this method as RoughDE. The proposed method is experimentally compared with other hybrid Rough Set methods such as Genetic Algorithm (GA) and Particle Swarm Optimization (PSO).

1 Introduction

The main goal of feature selection is to find a minimal feature subset from a problem domain with high accuracy in representing the original features [1]. In real world problems feature selection is an important process must due to the abundance of noisy, irrelevant or misleading features. An extensive method may be used for this purpose, it is quite impractical for most datasets. Usually feature selection algorithms involve heuristic or random search strategies in an attempt to avoid this prohibitive complexity. However, the degree of optimality of the final feature subset is often reduced.

Rough set theory provides a mathematical tool that can be used for both feature selection and knowledge discovery. It helps us to find out the minimal attribute sets called ‘reducts’ to classify objects without deterioration of classification quality. [2,3,4]. However, it is not possible in theory to say whether two attribute values are similar and to what extent they are the same; for example, two close values may only differ as a result of noise, but in the standard RST-based approach they are considered to be as different as two values of a different order of magnitude. Dataset discretization must take place before reduction methods based on crisp rough sets can be applied. This is often still inadequate. However, as the degrees of membership of values to discretised values are not considered at all. To solve this problem, a number of variations in this theory have been proposed. Among these methods, the Swarm Intelligence (SI) based methods perform better than the rest of the method.

Swarm Intelligence is the property of a system whereby the collective behaviours of simple agents interacting locally with their environment cause coherent functional global patterns to emerge [5]. SI provides a basis with which it is possible to explore collective (or distributed) problem solving without centralized control or the provision of a global model.

The rest of this paper is structured as follows. Section 2 discusses the fundamentals of Rough Set Theory, in particular focusing on dimensionality reduction. Section 3 presents the Concepts of Differential Evolution Optimization technique. Section 4 presents the hybrid methods of Rough Set theory with Differential Evolution (RoughDE). Section 5 details the experimentation carried out and presents the discovered results. The paper concludes with a discussion on the observations and highlights the scope for future work in this area.

2 Rough Set Preliminaries

Rough set theory [6][7] is a new mathematical approach to imprecision, vagueness and uncertainty. In an information system, every object of the universe is associated with some information. Objects characterized by the same information are indiscernible with respect to the available information about them. Any set of indiscernible objects is called an elementary set. Any union of elementary sets is referred to as a crisp set- otherwise a set is rough (imprecise, vague). Vague concepts cannot be characterized in terms of information about their elements. A rough set is the approximation of a vague concept by a pair of precise concepts, called lower and upper approximations. The lower approximation is a description of the domain objects, which are known with certainty to belong to the subset of interest, whereas the upper approximation is a description of the objects that possibly belong to the subset. Relative to a given set of attributes, a set is rough if its lower and upper approximations are not equal.

The main advantage of rough set analysis is that it requires no additional knowledge except for the supplied data. Rough sets perform feature selection using only the granularity structure of the data [5]. Let $I = (U, A)$ be an information system, where U is the universe, a non-empty finite set of objects. A is a non-empty finite set of attributes. For $\forall a \in A$ determines a function $f_a: U \rightarrow V_a$. If $P \subseteq A$, there is an associated equivalence relation:

$$IND(P) = \{(x, y) \in U \times U \mid \forall a \in P, f_a(x) = f_a(y)\} \tag{1}$$

The partition of U , generated by $IND(P)$ is denoted U/P . If $(x, y) \in IND(P)$, then x and y are indiscernible by attributes from P . The equivalence classes of the P -indiscernibility relation are denoted $[x]_P$. The indiscernibility relation is the mathematical basis of rough set theory.

Let $X \subseteq U$, the P -lower approximation \underline{PX} and P -upper approximation \overline{PX} of set X can be defined as:

$$\underline{PX} = \{ x \in U \mid [x]_P \subseteq X \} \tag{2}$$

$$\overline{PX} = \{ x \in U \mid [x]_P \cap X \neq \emptyset \} \tag{3}$$

Let $P, Q \subseteq A$ be equivalence relations over U , then the positive, negative and boundary regions can be defined as:

$$POS_P(Q) = \bigcup_{X \in U/Q} \frac{PX}{|PX|} \tag{4}$$

$$NES_P(Q) = U - \bigcup_{X \in U/Q} \overline{PX} \tag{5}$$

$$BND_P(Q) = \bigcup_{X \in U/Q} (\overline{PX} - \frac{PX}{|PX|}) \tag{6}$$

The positive region of the partition U/Q with respect to P , $POS_P(Q)$, is the set of all objects of U that can be certainly classified to blocks of the partition U/Q by means of P . A set is rough (imprecise) if it has a non-empty boundary region.

An important issue in data analysis is discovering dependencies between attributes. Dependency can be defined in the following way. For $P, Q \subseteq A$, P depends totally on Q , if and only if $IND(P) \subseteq IND(Q)$. That means that the partition generated by P is finer than the partition generated by Q . We say that Q depends on P in a degree $0 \leq k \leq 1$ denoted $P \Rightarrow_k Q$, if

$$K = \gamma_P(Q) = \frac{|POS_P(Q)|}{|U|} \tag{7}$$

If $k=1$, Q depends totally on P , if $0 \leq k < 1$, Q depends partially on P , and if $k=0$ then Q does not depend on P . In other words, Q depends totally (partially) on P , if all (some) objects of the universe U can be certainly classified to blocks of the partition U/Q , employing P .

In a decision system the attribute set contains the condition attribute set C and decision attribute set D , i.e. $A = C \cup D$. The degree of dependency between condition and decision attributes, $\gamma_C(Q)$, is called the quality of approximation of classification, induced by the set of decision attributes [6].

The goal of attribute reduction is to remove redundant attributes so that the reduced set provides the same quality of classification as the original. A reduct is defined as a subset R of the conditional attribute set C such that $\gamma_R(Q) = \gamma_C(Q)$. A given decision table may have many attribute reducts, the set of all reducts is defined as:

$$Red = \{R \subseteq C \mid \gamma_R(D) = \gamma_C(D), \forall B \subset R, \gamma_B(D) \neq \gamma_C(D)\} \tag{8}$$

In rough set attribute reduction, a reduct with minimal cardinality is searched for. An attempt is made to locate a single element of the minimal reduct set $Red_{min} \subseteq Red$:

$$Red_{min} = \{R \in Red \mid \forall R' \in Red, |R| \leq |R'| \} \tag{9}$$

The intersection of all reducts is called the core, the elements of which are those attributes that cannot be eliminated. The core is defined as:

$$CORE(C) = \bigcap Red \tag{10}$$

3 Differential Evolution (DE) and Its Modification

Differential evolution (DE) is well known as a simple and efficient scheme for global optimization over continuous spaces. The classical DE [8] is a population-based global optimization algorithm that uses a floating-point (real-coded) representation. The i th individual vector (chromosome) of the population at time-step (generation) t has d components (dimensions), i.e.

$$\vec{Z}_i(t) = [Z_{i,1}(t), Z_{i,2}(t), \dots, Z_{i,d}(t)] \tag{11}$$

For each individual vector $\vec{Z}_k(t)$ that belongs to the current population, DE randomly samples three other individuals, i.e., $\vec{Z}_i(t)$, $\vec{Z}_j(t)$ and $\vec{Z}_m(t)$, from the same generation (for distinct k, i, j , and m). It then calculates the (component wise) difference of $\vec{Z}_i(t)$ and $\vec{Z}_j(t)$, scales it by a scalar F (usually $\in [0,1]$), and creates a trial offspring $\vec{U}_i(t + 1)$ by adding the result to $\vec{Z}_m(t)$. Thus, for the n th component of each vector.

$$U_{k,n}(t+1) = \begin{cases} Z_{m,n}(t) + F(Z_{i,n}(t) - Z_{j,n}(t)), & \text{if } \text{rand}_n(0,1) < C_r \\ Z_{k,n}(t), & \text{otherwise} \end{cases} \tag{12}$$

$C_r \in [0,1]$ is a scalar parameter of the algorithm, called the crossover rate. If the new offspring yields a better value of the objective function, it replaces its parent in the next generation; otherwise the parent is retained in the population, i.e.,

$$\vec{Z}_i(t + 1) = \begin{cases} \vec{U}_i(t + 1), & \text{if } f(\vec{U}_i(t + 1)) > f(\vec{Z}_i(t)) \\ \vec{Z}_i(t) & \text{if } f(\vec{U}_i(t + 1)) \leq f(\vec{Z}_i(t)) \end{cases} \tag{13}$$

where $f(\cdot)$ is the objective function to be maximized.

To improve the convergence properties of DE, we have tuned its parameters in two different ways here. In the original DE, the difference vector $(\vec{Z}_i(t) - \vec{Z}_j(t))$ is scaled by a constant factor F . The usual choice for this control parameter is a number between 0.4 and 1. We propose to vary this scale factor in a random manner in the range (0.5, 1) by using the relation

$$F = 0.5 * (1 + \text{rand}(0, 1)) \tag{14}$$

where $\text{rand}(0, 1)$ is a uniformly distributed random number within the range [0, 1]. The mean value of the scale factor is 0.75. This allows for stochastic variations in the amplification of the difference vector and thus helps retain population diversity as the search progresses. The DE with random scale factor (DERANDSF) can meet or beat the classical DE. In addition to that, here we also linearly decrease the crossover rate C_r with time from $C_{r_{\max}} = 1.0$ to $C_{r_{\min}} = 0.5$. If $C_r = 1.0$, it means that all components of the parent vector are replaced by the difference vector operator. However, at the later stages of the optimizing process, if C_r is decreased, more components of the parent vector are then inherited by the offspring. Such a tuning of C_r helps exhaustively explore the search space at the beginning but finely adjust the movements of trial solutions during the later stages of search, so that they can explore

the interior of a relatively small space in which the suspected global optimum lies. The time variation of Cr may be expressed in the form of the following equation:

$$Cr = (Cr_{max} - Cr_{min}) * (MAXIT - iter)/MAXIT \tag{15}$$

where Cr_{max} and Cr_{min} are the maximum and minimum values of crossover rate Cr, respectively; iter is the current iteration number; and MAXIT is the maximum number of allowable iterations.

4 DE and Rough Set-Based Feature Selection (RoughDE)

We can use the idea of DE for the optimal feature selection problem. Consider a large feature space full of feature subsets. Each feature subset can be seen as a point or position in such a space. If there are N total features, then there will be 2^N kinds of subset, different from each other in the length and features contained in each subset. The optimal position is the subset with least length and highest classification quality. Now we put a particle swarm into this feature space, each particle takes one position.

To apply the DE idea to feature selection, some matters must first be considered.

Representation of position

We represent the particle's position as binary bit strings of length N, where N is the total number of attributes. Every bit represents an attribute, the value '1' means the corresponding attribute is selected while '0' not selected. Each position is an attribute subset.

Position Update Strategies

The number of different bits between two particles relates to the difference between their positions. For example, $Z_i(t)=[1\ 0\ 1\ 1\ 1\ 0\ 1\ 0\ 0\ 1]$, $Z_j(t)=[0\ 1\ 0\ 0\ 1\ 1\ 0\ 1\ 0\ 1]$. The difference between two particle is $Z_i(t) - Z_j(t)=[1\ -1\ 1\ 1\ 0\ -1\ 1\ -1\ 0\ 0]$. A value of 1 indicates that compared with the best position, this bit (feature) should be selected but is not, which will decrease classification quality and lead to a lower fitness value. Assume that the number of 1's is a . On the other hand, a value of -1 indicates that, compared with the best position, this bit should not be selected, but is selected. Redundant features will make the length of the subset longer and lead to a lower fitness value. The number of -1's is b . We use the value of $(a-b)$ to express the distance between two positions; $(a-b)$ may be positive or negative. Such variation makes particles exhibit an exploration ability within the solution space. In this example, $(a-b)=4-3=1$, so $Z_i(t) - Z_j(t) = 1$.

In our experimentation, we initially take 0 and 1 bit. However, it was noticed that after several generations the particles converged to a good but non-optimal solution. So particles will adjust positions, searching around the best position. After many tests, we found that the position of particle change to minimum and maximum value across the 0 and 1. So we update position of new generated particle as

$$U_{k,n}(t+1) = \begin{cases} 1, & \text{if } Z_{m,n}(t) + F(Z_{i,n}(t) - Z_{j,n}(t)) > 0 \\ 0, & \text{if } Z_{m,n}(t) + F(Z_{i,n}(t) - Z_{j,n}(t)) < 0 \end{cases} \tag{16}$$

Fitness Function

We define the fitness function in equation (17):

$$Fitness = \alpha * \gamma_R(D) + \beta * \left(\frac{|c| - |R|}{|R|} \right) \quad (17).$$

Where $\gamma_R(D)$ is the classification quality of condition attribute set R relative to decision D , $|R|$ is the '1' number of a position or the length of selected feature subset. $|C|$ is the total number of features. α and β are two parameters corresponding to the importance of classification quality and subset length, $\alpha \in [0,1]$ and $\beta = 1 - \alpha$.

This formula means that the classification quality and feature subset length have different significance for feature selection task. In our experiment we assume that classification quality is more important than subset length and set $\alpha = 0.9$, $\beta = 0.1$. The high α assures that the best position is at least a real rough set reduct. The goodness of each position is evaluated by this fitness function. The criteria are to maximize fitness values.

5 Experiment and Results

We run the proposed dimensionality reduction technique to reduce the dataset as lower dimensional dataset.

A. Experimental Setup

The parameters of the algorithms for all examples are defined as follows: the size of the population in GA, swarm size in PSO and particle size in DE are set to $(int)(10 + 2 * sqrt(D))$, where D is the dimension of the position, i.e., the number of condition attributes. Each experiment (for each algorithm) was repeated 3 times with different random seeds with number of **fitness function evaluation** is 600

B. Datasets Used

There are some real-life data sets are used in this paper which are taken from UCI Machine Repository and from website <http://www.ailab.si/orange/datasets.asp>.

C. Simulation Strategy

In this paper, while comparing the performance of our proposed **RoughDE** algorithm with other techniques, we focus on two major issues: as 1) ability to find the optimal number of attributes or features and 2) computational time required to find the solution. For comparing the speed of the algorithms, the first thing we require is a fair time measurement. The number of iterations or generations cannot be accepted as a time measure since the algorithms perform different amount of works in their inner loops, and they have different population sizes. Hence, we choose the number of **fitness function evaluations (FEs)** as a measure of computation time instead of generations or iterations. Since the algorithms are stochastic in nature, the results of two successive runs usually do not match. Hence, we have taken 3 independent runs (with different seeds of the random number generator) of each algorithm.

Finally, we would like to point out that all the experiment codes are implemented in MATLAB 7. The experiments are conducted on a Pentium 4, 1GB RAM, and the system is Windows XP Professional.

Table 1. Datasets used in the experiments

SL.No	Datasets	No. of Instances	No. of features	No. of clusters	RoughGA	RoughPSO	RoughDE
1	Iris	150	04	03	02-03	02-03	02
2	Glass	214	09	06	03	03	03
3	Wine	178	13	03	03	02-03	02-03
4	Vowel	462	10	11	3	02-04	02
5	Breast cancer wisconsin	683	9	2	4	4	2
6	Zoo	101	16	7	5	5	5
7	Pima Indian diabetes	768	8	2	2	2	2
8	shuttle landing control	253	6	2	6	6	4-6
9	Monk1	432	6	2	3	3	3
10	Lung cancer	32	56	3	21	21	19
11	Hayes roth	160	4	3	3	3	3
12	Ionosphere	351	32	2	12	12	12
13	Balance scale	625	4	2	4	4	2-4
14	Ecoli	336	7	8	3	3	3
15	Haberman's survival	306	3	2	3	3	3
16	Vechile	846	18	4	6	6	4

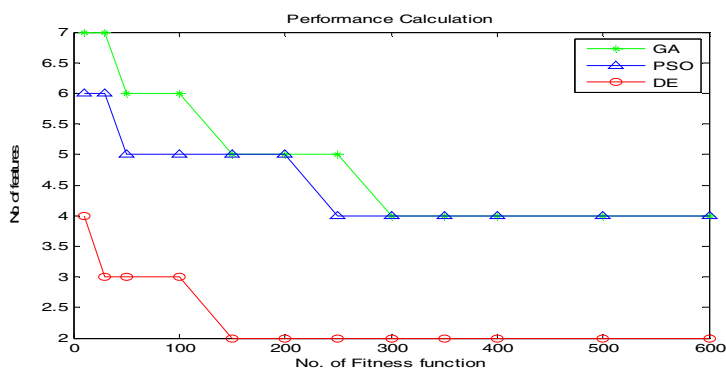


Fig. 1. Performance of rough set reduction for Wine dataset

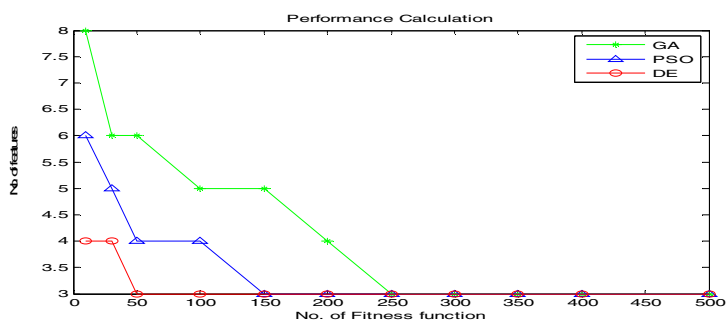


Fig. 2. Performance of rough set reduction for Breast cancer wisconsin dataset

D. Experimental Results

To judge the accuracy of the RoughDE and other algorithms, we let each of them run for a very long time over every benchmark data set, until the number of FEs exceeded 600. Then, we note the number of features found.

E. Discussion on Results

DE usually can obtain a better result than GA and PSO, especially for a large scale problem. Although GA, PSO and DE got the same results, DE usually uses only very few iterations, as illustrated in Fig. 1 and Fig. 2.

Conclusion

In this paper, we investigated the problem of finding optimal reducts using Differential evolution technique. The proposed approach discovered the best feature combinations in an efficient way to observe the change of positive region as the particles proceed throughout the search space. We evaluated the performance of the proposed RoughDE algorithm with Genetic Algorithm (GA) and RoughPSO. The results indicate that DE usually required shorter time to obtain better results than GA and PSO, specially for large scale problems, although its stability needs to be improved in further research. The proposed algorithm could be an ideal approach for solving the reduction problem when other algorithms failed to give a better solution.

References

1. Dash, M., Liu, H.: Feature Selection for Classification. *Intelligent Data Analysis* 1(3), 131–156 (1997)
2. Pawlak, Z.: Rough Sets. *International Journal of Computer and Information Sciences* 11, 341–356 (1982)
3. Pawlak, Z.: Rough Sets: Present State and The Future. *Foundations of Computing and Decision Sciences* 18, 157–166 (1993)
4. Pawlak, Z.: Rough Sets and Intelligent Data Analysis. *Information Sciences* 147, 1–12 (2002)
5. Bonabeau, E., Dorigo, M., Theraulez, G.: *Swarm Intelligence: From Natural to Artificial Systems*. Oxford University Press Inc., New York (1999)
6. Pawlak, Z.: *Rough Sets: Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishing, Dordrecht (1991)
7. Polkowski, L., Lin, T.Y., Tsumoto, S. (eds.): *Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems*. *Studies in Fuzziness and Soft Computing*, vol. 56. Physica-Verlag, Heidelberg (2000)
8. Storn, R., Price, K.: Differential evolution—A simple and efficient heuristic for global optimization over continuous spaces. *J. Glob. Optim.* 11(4), 341–359 (1997)
9. Thangavel, K., Shen, Q., Pethalakshmi, A.: Application of Clustering for Feature Selection Based on Rough Set Theory Approach. *AIML Journal* 6(1) (January 2006)
10. Yue, B., Yao, W., Abraham, A., Liu, H.: A New Rough Set Reduct Algorithm Based on Particle Swarm Optimization. In: Mira, J., Álvarez, J.R. (eds.) *IWINAC 2007*. LNCS, vol. 4527, pp. 397–406. Springer, Heidelberg (2007)
11. Jeba Emilyn, J., Ramar, K.: Rough Set Based Clustering Of Gene Expression Data: A Survey. *International Journal of Engineering Science and Technology* 2(12), 7160–7164 (2010)

An Enhanced Scheduling Strategy to Accelerate the Business Performance of the Cloud System

T.R. Gopalakrishnan Nair¹, M. Vaidehi², K.S. Rashmi³, and V. Suma²

¹ Research, Dayananda Sagar Institutions,
ARAMCO Endowed Chair –Technology, PMU, KSA
trgnair@ieee.org

² Research and Industry Incubation Center,
Dayananda Sagar Institutions, Bangalore, India
{dm.vaidehi, sumavdsce}@gmail.com

³ Post Graduate Programme, Computer Science and Engineering,
Department of Information Science and Engineering,
Dayananda Sagar College of Engineering, Bangalore
rashmiks.ks57@gmail.com

Abstract. The prime objective of any IT organization is to develop cost, time and resource effective products. In any organization, jobs arriving to the system at peak hours are normally high demanding efficient execution and dispatch of jobs. Cloud computing is an emerging technology, which enables one to accomplish aforementioned objective, leading towards improved business performance. The observations made by capturing a job arriving pattern from monitoring system indicates most of the jobs are rejected due to inefficient scheduling technique. This paper therefore introduces an enhanced scheduling strategy that incorporates prioritization of jobs based on criticality and business gain in the Round Robin task scheduling technique. Simulation results have shown an improved response time and processing time of jobs. Further, the number of jobs that get rejected has decreased despite of geographically dispersed users, datacenters or the processing units and thereby increasing the number of computation of tasks.

Keywords: Cloud Computing, Virtualization, Enhanced Round Robin, Scheduling, Response time, Processing time.

1 Introduction

Evolution of IT industry has oriented towards the consumption of large scale infrastructure, advanced operating platforms and development of optimal software and thereby demanding heavy capital investment by the organization. The state-of-art of the technology focuses on data processing to deal with massive amount of data. Thus, data processing has emerged as one of the challenging areas of any organization.

Cloud computing is a promising technology, which is a pay-go model that provides the required resources to its clients. The main characteristics of cloud computing includes support of client-server model, distributed computing, utility computing and

virtualization. Since, virtualization is one of the core characteristics of cloud computing it is now possible to virtualize the factors that modulates business performance such as IT resources, hardware, software and operating system in the cloud-computing platform. Additionally, the diverse application of various users runs independently on the virtualized operating system. Consequently, resources are sacrificed in order to perform the activities on individual unit of services and depriving the execution of other activities, which are requesting for the same resource. Hence, an efficient task scheduling is an imperative factor in the cloud environment, failure of which leads to performance degradation of the cloud system. An efficient task scheduling aim towards less response time and henceforth the submitted jobs will be processed within the stipulated period based on their size. Subsequently, the resources are reallocated in time and enable to submit more number of jobs to the cloud. This in turn accelerates the business performance of cloud system.

This paper introduces an efficient task scheduling strategy where Enhanced Round Robin (ERR) scheduling technique is incorporated in VM scheduler to achieve above-mentioned objectives. This technique in turn is compared with the current Round Robin (RR) scheduling technique.

2 Research Background

The advancement of technology in IT domain has opened several research avenues in order to enhance the business performance. Shou Liu et al. have proposed an accrual scheduling algorithm for real-time cloud computing services to improve the performance of cloud over the traditional utility accrued approach [1].

Authors in [2] have suggested a Cost-optimal scheduling technique in hybrid IaaS clouds for deadline-constrained workloads, which aims to maximize the utilization of the internal datacenter and to minimize the cost of running outsourced tasks in the cloud within the specified QoS constraints.

Further, authors in [3] have recommended a three-phases scheduling load-balancing algorithm in hierarchical cloud computing network. They suggest selecting the service nodes on the network to facilitate the execution of complicated tasks that requires large-scale computations.

However, Ana-Maria Operscu et al. have implemented Bag-of-Tasks Scheduling (BaTS) technique to schedule large bags of tasks on the multiple clouds having different CPU performance and cost. The authors further feel that the implementation of their technique minimizes the computation time without the need of task completion time [4].

Authors in [5] have incorporated an improved cost-based scheduling algorithm for task scheduling in cloud computing with the objective to schedule task groups in cloud computing platform having resources with different costs and computation performance.

Authors in [6] have suggested a task scheduling technique using credit based assignment problem in cloud computing to find the minimal completion time of tasks.

Nevertheless, the progress of research in cloud environment, it is apparent that currently existing job scheduling techniques are inefficient in reducing the job

rejection at peak hours, which is observed in the day-to-day scenario. Hence, this research focuses towards enhancing the existing task scheduling strategy in order to increase the computation of more number of jobs and thus improving the business performance of the organization.

3 Research Design

An efficient resource model enhances the performance of computing system in cloud environment. Since, cloud is a pay-and-go model, an unplanned scheduling of tasks leads to degradation of system and business performance. Therefore, it is essential for an efficient resource model to support optimized utilization of available resources and execution of tasks within the stipulated time in the cloud environment.

Data collection for this research includes secondary data in order to analyze the efficiency of cloud service provider for the successful execution of jobs at the peak hours and within the available resource and time. Secondary data are processed data collected from several leading IT organizations, which are operating in cloud environment.

It is apparent from the analysis that most of the jobs are rejected due to inefficient scheduling technique. Hence, the aim of this research was to develop an enhanced task scheduling strategy, which can resolve the aforementioned issues.

4 Research Method

The existing task scheduling techniques operated for cloud in IT organizations include Round Robin (RR) scheduling, BaTS, ABC scheduling, three-phases scheduling etc. However, our deep investigations have evidently shown the inefficiency of existing task scheduling techniques in terms of response time and processing time. Consequently, the rejection ratio of jobs is quite high.

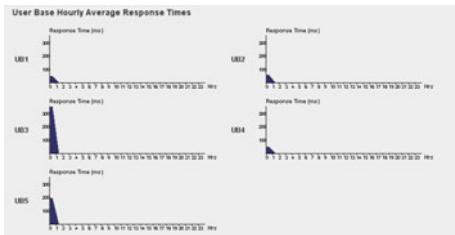


Fig. 1. A Graph of User Base Hourly Average Response time using RR scheduler



Fig. 2. Graph Showing the number of jobs rejected using RR

Fig. 1. shows a sample graph of user base hourly average response time using the RR scheduler. From the graph it is analysed that there is a high response time leading to very high processing time and ultimately resulting in increased rejection in the

number of the jobs submitted. Fig. 2. illustrates the number of jobs rejected at the peak hours. This figure further infers that with increase in the submission rate of jobs, rejection rate further increases.

Our Enhanced Round Robin (ERR) task scheduling technique incorporates prioritization of jobs based on task criticality and business gain in the Round Robin (RR) task scheduling technique. Fig. 3. shows the design of our model to implement ERR task scheduling technique. Accordingly, various jobs are submitted from diverse clients to the cloud service provider via a communication channel. The task manager prioritizes the queue of jobs based on task criticality and business gains [9]. Subsequently the priority scheduler schedules the jobs from prioritized queue to the virtual machine (VM).

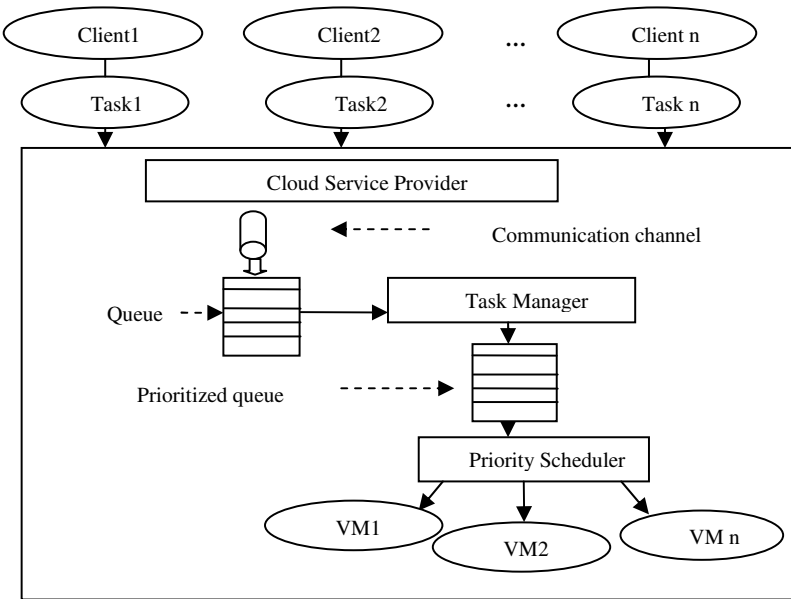


Fig. 3. Cloud Architecture

5 Algorithm to Implement ERR

Algorithm which is used to implement the ERR model is given below.

Algorithm: Enhanced Round Robin (ERR) Task Scheduler.

1. ERR scheduler maintains an index table of the VMs with the state (Busy/Available) and the priorities of the tasks (High, Medium and Low).
2. As the request is received by the task manager, it queries the ERR scheduler for the next allocation.
3. ERR scheduler allocates the next available VM based on the priority of the task.

4. If the VM is available ERR scheduler assigns that VM to the task with the highest priority.
5. If not the request will be queued.
6. Task manager checks for the waiting requests in the queue. If there are it continues from step3.

6 Simulation Setup

For the implementation of the above algorithm, the simulation tool namely Cloud Analyst is used. The simulation setup for ERR scheduler comprises of user base configuration, datacenter configuration and advanced configuration. Table 1. depicts the details of user base configuration. Table 2. and Table 3. shows the information of datacenter configuration and advanced configuration respectively.

The user base configuration specifies details related to request for every hour, data size for every request and datacenter where the requests are processed.

Table 1. User Base Configuration

UB	R/U/H	DS/R	DC
UB1	12	100	DC4
UB2	12	10000	DC3
UB3	12	100000	DC1
UB4	12	1000	DC2
UB5	12	10000	DC2

Table 2. Datacenter Configuration

DC	No. of VMs	Mem.	BW	CPV \$/Hr	Mc \$/s	Sc \$/s	Dc \$/Gb
DC1	40	1024	1000	0.1	0.05	0.1	0.1
DC2	20	512	100	0.1	0.05	0.1	0.1
DC3	50	512	10000	0.1	0.05	0.1	0.1
DC4	35	1024	1000	0.1	0.05	0.1	0.1

DC – Datacenter, Mem. – Memory, BW –Bandwidth, Mc–Memory cost, Sc – Storage cost, Dc- Data Transfer cost.

UB – User base,
 R/U/H –Request per user per hour, DS/R
 – Data Size per request,
 DC – Datacenter.

Table 3. Advanced Configuration

User grouping factor	1000
Request grouping factor	100
Executable instruction length	250

Datacenter configuration provides information such as number of VMs present in DC, memory and bandwidth capability of every VM, memory, storage and data transfer cost of every VM per hour.

The advanced configuration of cloud analyst tool comprises of user grouping factor in user base, request grouping factor in datacenters and executable instruction length for every request.

7 Results

Simulation of ERR scheduler has yielded appreciable results in terms of response time and processing time and henceforth resulting in reduced rejection of submitted jobs. Table 4. depicts comparative analysis of response time and processing time with respect to RR scheduling and ERR scheduling. The table further infers that ERR scheduler provides better results than compared to RR scheduler.

Fig. 1. and Fig. 4. indicates the sample graphs of user base hourly average response time using the RR scheduler and the ERR scheduler respectively. The comparison of the 2 figures strongly indicates that the less response time of jobs using ERR scheduler results in less processing time of the jobs.

Table 5. depicts a comparative analysis of RR scheduler and ERR scheduler in terms of response time with ten different configurations. Inferences from the table indicates that the response time for processing the jobs using ERR scheduler has reduced up to 30% when compared to the response time for processing the jobs using

Table 4. Comparative analysis with respect to Response Time and Processing Time

Mm	RT (ms)		DC PT (ms)	
Scheduler	RR	ERR	RR	ERR
Avg. (ms)	115.23	84.93	2.51	1.87
Min. (ms)	44.33	32.68	0.08	0.06
Max. (ms)	380.03	279.9	7.98	5.95

Mm – Measurement metric, RT –Response time, DC PT – Datacenter Processing Time.

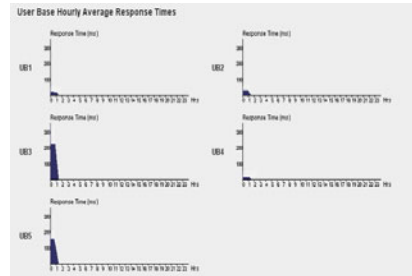


Fig. 4. A Graph of User Base Hourly Average Response time using ERR Scheduler

Table 5. A Comparison of RR scheduler and ERR scheduler in terms of Response Time

Request per user per hour	Response Time (ms)		Percentage Variation (%)
	RR scheduler	ERR scheduler	
12	115.23	84.93	26.29
24	118.16	84.37	28.6
45	114.87	85.35	25.7
35	118.33	83.01	29.85
60	115.80	84.3	27.24
30	109.94	82.34	25.12
70	113.97	81.92	28.12
75	109.90	77.81	29.2
80	120.44	88.72	26.34
90	123.10	88.98	27.72

Table 6. A Comparison of RR scheduler and ERR scheduler in terms of Datacenter Processing time

Request per user per hour	Datacenter Processing Time (ms)		Percentage Variation (%)
	RR scheduler	ERR scheduler	
12	2.51	1.87	25.49
24	2.41	1.76	27.2
45	2.36	1.7	28.08
35	2.31	1.71	26.01
60	2.36	1.72	27.1
30	2.37	1.74	26.61
70	2.36	1.76	25.34
75	2.43	1.75	27.9
80	2.43	1.77	27.26
90	2.42	1.73	28.37

RR scheduler. Similarly, Table 6. Depicts a comparative result of RR scheduler and ERR scheduler in terms of datacenter processing time with ten different configurations. The observations from the table indicates that the datacenter processing time of using ERR scheduler has reduced up to 30% when compared to the datacenter processing time of jobs by using RR scheduler.

At peak hour since the number of jobs arriving to the system is high and, if the number of jobs rejected increases, the system undergoes performance degradation. Fig. 5. depicts the rejected number of submitted jobs using ERR scheduler. From the simulation it is observed that due to the implementation of ERR scheduler the submitted jobs require less response time and less processing time and thus facilitates more number of jobs to be submitted and processed.

Table 7. and Fig. 6. shows a comparison between RR scheduler and ERR scheduler in terms of percentage of rejection of the submitted jobs. It is apparent from these table and figure that ERR scheduler gives less percentage of rejection of submitted jobs. Consequently, the application of ERR scheduler enhances the business gain. However, the future work is to evaluate the performance of ERR scheduler with other existing standard metrics.

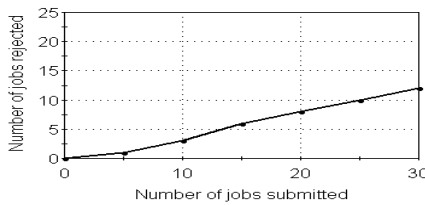


Fig. 5. Graph showing less number of job rejections using the ERR scheduler

Table 7. Comparison of Rejected Number of submitted jobs using RR scheduler and ERR scheduler

No. of JS	JRR (%)	JERR (%)
5	40	30
10	40	30
15	53	40
20	75	40
25	76	40
30	80	40

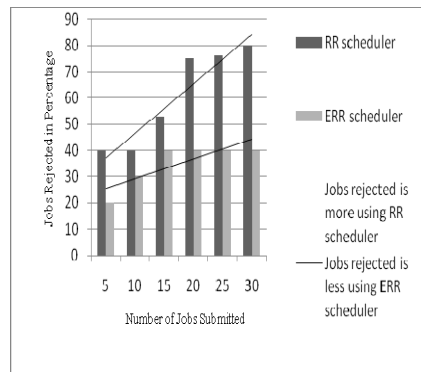


Fig. 6. A graph showing of comparison between RR scheduler and ERR scheduler in terms of percentage of jobs rejected

JS - Jobs Submitted,
 JRR - Jobs rejected using RR scheduler,
 JERR - Jobs rejected using ERR scheduler.

8 Conclusion

Advancement of technology in the IT industries requires development of optimal software in addition to consumption of large scale infrastructure, advanced operating platforms and heavy capital investment by the organization. Cloud computing is an emerging technology, which enables to achieve the above specified objective.

Number of jobs to be processed and executed by the system at peak hours is normally high. However, current job scheduling techniques in cloud environment are inefficient in reducing the number of jobs that get rejected at peak hours.

This paper introduces an enhanced scheduling technique that integrates prioritization of jobs based on task criticality and business gain in the Round Robin (RR) scheduling technique. Simulation of Enhanced Round Robin (ERR) scheduler has yielded appreciable results in terms of response time and processing time when compared to the RR scheduler and henceforth resulting in reduction of rejection of submitted jobs. As less rejection of submitted jobs has been achieved using the ERR scheduler there will be an improvement in the business performance.

References

1. Liu, S., Quan, G., Ren, S.: On-line Scheduling of Real-time Services for Cloud Computing. In: 2010 IEEE 6th World Congress on Services, pp. 459–464. IEEE Computer Society (2010)
2. Van den Bossche, R., Vanmechelen, K., Broeckhove, J.: Cost-Optimal Scheduling in Hybrid IaaS Clouds for Deadline Constrained Workloads. In: 2010 IEEE 3rd International Conference on Cloud Computing, pp. 228–235. IEEE Computer Society (2010)
3. Wang, S.-C., Yan, K.-Q., Wang, S.-S., Chen, C.-W.: A Three-Phases Scheduling in a Hierarchical Cloud Computing Network. In: 2011 Third International Conference on Communications and Mobile Computing, pp. 114–117. IEEE Computer Society (2011)
4. Oprescu, A., Kielmann, T.: Bag-of-Tasks Scheduling under Budget Constraints. In: 2010 2nd IEEE International Conference on Cloud Computing Technology and Science, pp. 351–359. IEEE Computer Society (2010)
5. Selvarani, S., Sadhasivam, G.S.: Improved Cost-Based Algorithm for Task Scheduling in Cloud Computing. In: 2010 IEEE, pp. 1–5 (2010)
6. Paul, M., Sanyal, G.: Task-Scheduling in Cloud Computing using Credit Based Assignment Problem. *International Journal on Computer Science and Engineering (IJCSSE)* 3(10), 3426–3430 (2011)
7. Gong, C., Liu, J., Zhang, Q., Chen, H., Gong, Z.: The characteristics of cloud computing. In: 2010 IEEE International Conference on Parallel Processing Workshops, pp. 275–279 (2010)
8. Wang, X., Wang, B., Huang, J.: Cloud computing and its key techniques. In: 2011 IEEE International Conference on Computer Science and Automation Engineering, pp. 404–410 (2011)
9. Nair, T.R.G., Vaidehi, M.: Efficient resource arbitration and allocation strategies in cloud computing through virtualization. In: 2011 IEEE International Conference on Cloud Computing and Intelligence Systems, Beijing, China, pp. 397–401 (September 2011)
10. Zhang, S., Chen, X., Zhang, S., Huo, X.: The comparison between cloud computing and grid computing. In: 2010 IEEE International Conference on Computer Application and System Modeling (ICCASM), pp. V11-72–V11-75 (2010)

A Novel Approach for Intrusion Detection Using Swarm Intelligence

M. Sailaja¹, R. Kiran Kumar², P. Sita Rama Murty³,
and P.E.S.N. Krishna Prasad⁴

¹ Dept. of ECE, JNTUK University, Kakinada, India

² Department of CSE, Krishna University, Machilipatnam, India

³ Department of IT, Sri Sai Aditya Institute of Science and Technology,
Kakinada, India

⁴ Dept. of CSE, Aditya Engineering College, Kakinada, India
{s.maruvada}@rediffmail.com,
{kirankreddi,psramam.1,surya125}@gmail.com

Abstract. This paper presents Architecture (IHDAIDS) for Intrusion Detection in wired networks which used Hybrid Intelligent Intrusion Detection Techniques making use of Swarm Intelligence algorithms. The architecture proposed in this paper is Intelligent, Hybrid, Distributed and Adaptive which makes real-time and dependable decisions regarding the intrusions; the Intelligent and Adaptive nature of the engine minimizes the rate of false positives and increases the accuracy, also reduces human intervention. This architecture is also concentrated on the data collection module as the quality of the intrusion detection depends on the data provided to the intrusion detection engine. We used a Hybrid Swarm Intelligence algorithm PSOACO2 (Particle Swarm Optimization/Ant Colony Optimization) for intrusion detection and compared the results with SVM (Support Vector Machine). For experiments we considered KDDCUP'99 Data which is widely used by intrusion detection researchers as a standard.

Keywords: Intrusion Detection Systems (IDS), Intrusion Detection Architecture, SVM, LibSVM, PSO/ACO2, DE.

1 Introduction

Information Systems become more powerful and are everywhere today; information systems of different organizations especially financial organizations are the target for many attackers. Traditional security methods like user authentication through password, encryption of confidential data, strictly framing access rights, and firewalls used at the perimeter of the network are used as first line of defense for network security. If a password is compromised, password authentication cannot prevent the access of the network resources by unauthorized users, even firewalls will fail if there is an error in configuration i.e., these security measures are unable to protect against the insider attacks and unsecured modems. Case studies have

shown that a vast majority of security attacks originate from the inside personnel of the organization. In fact some studies report that as much as 75% of all attacks are from someone with in an organization or someone with inside information [4]. This makes it important to monitor what is taking place on a system and look for suspicious behavior. Intrusion Detection Systems just do that. Intrusion is any set of actions performed by user, process, or software which attempt to compromise the security (integrity/confidentiality/availability) of a resource. Traditionally Intrusion Prevention is done by firewalls and encryption software; it provides protection for the network from external attacks. To obtain acceptable of security these traditional security solutions should be coupled with Intrusion Detection Systems. Intrusion Detection System continuously monitors the network (Intranet) when a possible intrusion is detected it will alert the users.

We proposed a framework for intrusion detection, which includes different modules like data collection, intelligent intrusion detection and active response by the intrusion detection system. Whenever a novel attack is found the signature of that attack is immediately updated in the knowledge repository of the intrusion detection engine by the feedback mechanism. We considered two approaches in the first approach we used SVM for Intelligent Intrusion Detection System and in the second approach PSO/ACO2 is considered to detect anomaly.

The rest of the paper is organized as follows, related work in section 2 where the different types of Intrusion Detection Systems/architectures are short discussed, followed by the proposed framework in section 3, and experimental setup in section 4 where we gave an introduction to the datasets used for these experiments and data pre-processing methods are described. Experimental results along with conclusion and future work are in section 5.

2 Related Work

Intrusion detection system is just like a burglar-alarm, which rises alarm when a burglar tires to trespass in to our compound. Intrusion detection systems are software or hardware systems that monitor the events occurring in computer systems or network and rises alarms whenever an intrusion is suspected, which will alert the network administrator.

1. With quick detection of an intrusion, the intruder can be identified and ejected from the network/system before any damage is done or any data is compromised.
2. IDSs also serve as a deterrent to intrusions.
3. Intrusion detection enables the collection of information about intrusion techniques that can be helped to strengthen the intrusion prevention techniques.

IDSs are based on the assumption that the behaviour of an intruder differs from the behaviour of a legitimate user. Patterns of legitimate user behaviour are established by observing past history, and if any significant deviations from these patterns is detected in the network, IDS will generate an alarm. Sometimes

there may be an overlap in the behaviour of the authorised user and the intruder, and this may lead to false positives (or false negatives), the performance of IDS is measured based on the rate of false positives and false negatives.

2.1 Intrusion Detection Methods

Depending on the techniques used, IDS can be classified as follows

Signature (or misuse) based ID: In this method, the data captured by the intrusion detection sensors is compared with the pre-known attack patterns (signatures) if any similarity is observed, then it is treated as intrusion. To employ this method we should possess the attack patterns (signatures) with us. This approach fails in case of new attacks. The signature base should be continuously updated, i.e., whenever a new attack is observed it should be updated in the rule base.

Anomaly based ID: This is based on the assumption that the behaviour of an intruder differs from the legitimate user behaviour in a quantifiable way. From the historical data normal expected behaviour / profiles of systems and users is constructed. This approach frequently produces false alarms as the behaviour of the users varies widely.

Many intrusion detection systems employ both anomaly and signature based intrusion detection methods.

2.2 Intrusion Detection Systems

There are two types of intrusion detection systems that employ the above mentioned intrusion detection methods. They are Network based IDS and Host Based IDS.

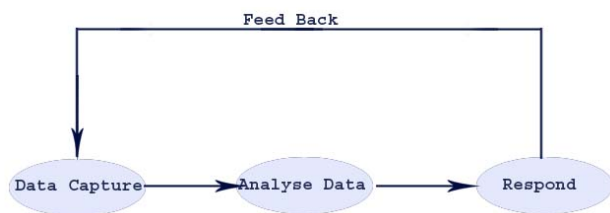


Fig. 1. Simple IDS Architecture

Network Intrusion Detection Systems (NIDS) monitors packets on the network wire and attempts to discover intrusion. NIDS may run either on the target machine who watches its own traffic or on an independent machine promiscuously watching all network traffic (hub, router, and probe).

Host based Intrusion Detection Systems base their detections based on the information obtained from a single host. In addition, host based IDS are able

to monitor accesses and changes to critical system files and changes in user privilege.

Both Network based IDS and Host based IDS have their pro's and con's, to achieve better results it is good to use combination of both.

3 Proposed Model

An ideal Intrusion Detection System should be Intelligent, adaptable, cost effective and capable of detecting intrusions in real time. We propose an Intelligent Hybrid, Distributed and Active Intrusion Detection System which collects data from multiple sources.

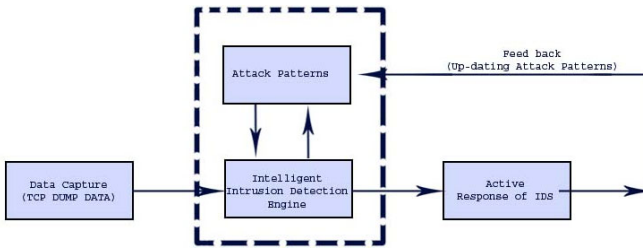


Fig. 2. IHDAIDS Architecture

IHDAIDS is characterised by the features like,

1. Its intelligent detection engine employing Swarm Intelligence algorithms, which to the maximum extent able to identify new attacks also.
2. Its hybrid nature, i.e., incorporating both Signature-based and Anomaly based intrusion detection methods.
3. Distributed architecture monitoring individual hosts and the network traffic.
4. Intrusion Detection System should provide active responses, as attacks are short lived and need an immediate attention, which is not practically possible for a human system administrator, if he have to do it himself.

In the present paper we concentrated on the Intelligent Intrusion Detection Engine of the IHDAIDS. Here we came up with two approaches, in first approach we considered Support Vector Machine and use LibSVM as the SVM tool, and in second approach we considered using a Hybrid PSO/ACO2 algorithm. The results of comparative study and performance analysis of these two algorithms when applied to intrusion detection is presented in section 5.

3.1 Data Collection Mechanism

Data collection systems for Intrusions Detection Systems can be classified as Host based data collection and Network based data collection [6].

Host based data collection allows collection of data that reflects accurately what is happening on the host, and this is direct monitoring i.e., we can obtain the data directly from the object that generates it. Most of the Intrusions that existing intrusion detection systems can detect are caused by actions performed in host.

Network based data collection is done by monitoring the traffic on the network, like capturing packets on the network using software centric techniques that rely on standard network interface cards (NICs). This technique is relatively fine at relatively slow speeds (2Gb/sec), but as the line speeds increase relativity and performance of these systems decrease. These are subjected to insertion and evasion attacks. In the above mentioned types of data capture Host based data capture proved to yield more accurate responses compared to Network based data capture because of lack of 100% packet capture capabilities. We need to capture, with absolute certainty, 100% of packets from wire, regardless of the network interface or packet rate, if we achieve this then automatically the intrusion detection becomes more accurate reducing false positives and false negatives.

The study of data collection mechanisms is important because "If the primary input (data collected) is incomplete or of low quality, how intelligent the intrusion detection engine may be, its results cannot be relied".

To get complete and qualitative data security events must be analyzed from as many sources as possible in order to asses threat and formulate appropriate response.

We propose a single authoritative source of real-time data that can be used to feed all monitoring and security applications, at the same time.

3.2 Intelligent Intrusion Detection Engine

Intelligent IDS apply soft computing methods for Intrusion Detection by using various Artificial Intelligence techniques to automate Intrusion Detection Process which reduce human intervention. Several artificial intelligence paradigms like neural networks, linear genetic programming, support vector machines, Bayesian networks, Fuzzy inference systems, etc., have been investigated for the design of IDS.

Support Vector Machines (SVMs) are being increasingly applied for intrusion detection since last decade. SVMs suit for intrusion detection because of their quick and effective learning from higher dimension data.

An SVM is a supervised learning algorithm that non-linearly maps input feature vectors into a higher dimensional feature space labelling each vector by its class. SVMs are developed based on the principal of structural risk minimization. Structural risk minimization seeks to find a hypothesis h for which one can find lowest probability of error where as the traditional learning techniques of pattern recognition is based on the minimization of the empirical risk, which attempt to optimize the performance of the learning set. Computing the hyper plane to separate the data points i.e. training an SVM leads to quadratic optimization problem. SVM uses linear separating hyper plane to create a classifier but all

the problems cannot be separated linearly in the original input space. SVM user a feature called kernel to solve this problem. The kernel transforms linear algorithms into non-linear ones via a map into feature spaces. There are many kernel functions including polynomial, radial basis functions, two layer sigmoid neural nets etc. The user may provide one of these functions at the time of training the classifier, which selects support vectors along the surface of this function. SVMs classify data by using support vectors, which are members of the set of training inputs that outline a hyper plane in feature space. We consider SVM for Intrusion detection because of their speed [11].

Particle Swarm Optimization/Ant Colony Optimization (PSO/ACO2) is proposed by N. Holden, and A.A. Freitas. This hybrid algorithm is capable of handling both nominal attribute values and continuous data values.

```

Initialize Population
Repeat for Max.Interactions
For every particle p
/* Rule Creation */
Set Rule R="If emptyset THEN c"
  For every dimension d in p
  Use roulette selection to choose whether the state should
  be a set to off or on.
  If it is on then the corresponding attribute-value pair
      set in the initialization will be added to R
  otherwise, if off is selected nothing will be added.
  Loop
    Calculate Quality Q of R
    /* Set the past best position */
    If Q>p's best part rule's (RP) Quaity Qp
      Qp=Q
      Rp=R
    END IF
  LOOP FOR every particle P
  Find best neighbour particle N according to N's Qp for every
  dimension d in P.
  /* Pheromone Updating Procedure */
  If best state is selected for Pd = best state selected for Nd then
  Pheromone_entry for the best state selected for Pd is increased by Qp.
  ELSE
  Pheromone_entry for the best state selected for Pd is decreased by Q.
  END IF
  Normalize Pheromone_entries
  LOOP
  LOOP
  LOOP

```

Pseudo-code for PSO/ACO2 [34]

PSO/ACO2 is mainly used to discover classification rules in the context of data mining. The knowledge or patterns discovered in the data set can be represented in terms of set of rules. A rule consists of antecedent (a set of attribute-values) and consequent (class).

Rule: IF $\langle \text{attrib} = \text{value} \rangle$ AND. $\langle \text{attribute}, \text{operator}, \text{value} \rangle$. AND
 $\langle \text{attrib} = \text{value} \rangle$ THEN $\langle \text{class} \rangle$

$\langle \text{attribute}, \text{operator}, \text{value} \rangle$ is the general term in the rule. For nominal attributes operator used is ' $=$ ', for continuous attributes ' $<=$ ' or ' $<$ ' is used.

PSO/ACO2 uses sequential covering approach, which discover one-rule at one-time. Here we have briefly explained about PSO/ACO2 the interested readers can find information about this algorithm in [34].

Differential Evolution Algorithm (DE) is a latest evolutionary optimization technique, which is population based, powerful and robust. DE starts with initializing the population; each individual in the population is an m-dimensional vector with random and uniform parameter values which lies in the predefined search space. For mutation without using probability distribution in DE, the weighted difference between two randomly selected individuals is added to a third individual generating mutation solution; this makes DE self organizing [12,13].

3.3 IDS Response (Active)

Active IDS is a system which automatically blocks suspected attacks in progression without any intervention required by operators, i.e., it can run a script to turn on or off a process, modify file permissions, terminate the offending processes, log of specific users etc. This system can send a TCP reset message to tell both sides of the connection to drop the session and stop communicating immediately.

3.4 Feedback

Security is a process not a product. How secure our system be it is for sure novel attacks will come into picture which may sometimes able to break into our so secure system. Whenever a security breach occurs in our system, immediately signature of that attack is added in the attack class's database of the Intrusion Detection Engine, such that if such attack is attempted again it can be detected immediately.

4 Experimental Setup

In this section we summarise our experimental results to detect intrusions using SVM and PSO/ACO2. We took the KDD'99 dataset to do our experiment. All the experiments were performed on Intel(R) Core(TM) i3 CPU, running at 2.53 GHz. The system has 3GB of RAM and Windows XP Operating System.

4.1 Dataset and Pre-processing

under the sponsorship of Defence Advanced Research Projects Agency(DARPA) and Air Force research Laboratory(AFRL) MIT Lincoln Laboratory has collected and distributed the datasets for the evolution of computer intrusion detection systems. The DARPA dataset is the most popular dataset used to test and evaluate large number of IDSs. The KDD'99 [5] is the subset of DARPA dataset. The dataset was pre-processed by extracting 41 features from the tcp-dump data from the 1998 DARPA dataset.

The DARPA dataset is about 4GB of compressed binary data, which is network traffic collected for 7 weeks. The full training set, one of the KDDCUP'99 datasets has 4,898,431 connections, which contains attacks. The attacks in the dataset fall into four categories

1. **DOS:** denial of service, e.g. syn_flood
2. **R2L:** unauthorized access from a remote machine, e.g. guessing password
3. **U2R:** unauthorized access to local super user (root) privileges, e.g. various "buffer overflow" attacks
4. **Probing:** surveillance and other probing, e.g. port scanning

KDD'99 features can be classified into three groups:

Basic Features: This category encapsulates all the attributes that can be extracted from a TCP/IP connection. Most of these features lead to implicit delay in detection.

Traffic Features: This category includes features that are computed with respect to a window interval. These are time based.

Content Features: we need some features to be able to look for suspicious behaviour in the data portion e.g. number of failed login attempts.

KDDCUP'99 data is available as a full training set, a 10% version of this training set, and a test set.

For the experiments we have constructed three samples each with 6000 records selected randomly from the Standard KDDCUP'99 Intrusion Detection Dataset. These are 10%_sample, 20%_sample and 25%_sample. 10%_sample contains 10% of attack records, similarly 20%_sample contains 20% of attacks and 25%_sample contains 25% of attacks.

4.2 Performance Measure

The research in IDS focuses on either to minimize the false alarms rate or to maximize detection rate (the rate of attacks detected successfully). High detection rate and low false positive rate are required for any good intrusion detection system.

False Positive: Incorrectly classifying normal data as an intrusion i.e. raising a alarm without any intrusion.

$$FalsePositiveRate = \frac{\#Normal_data_classified_as_Intrusions}{\#Intrusions} \tag{1}$$

False Negative: Incorrectly classifying an intrusion as normal.

$$FalseNegativeRate = \frac{\#Intrusions_as_Normal}{\#Intrusions} \tag{2}$$

True Positive: Identifying an intrusion as intrusion. We want to maximize True Positive rate. True positive rate is also known as sensitivity.

$$TruePositiveRate = \frac{\#Correct_Intrusions}{\#Intrusions} \tag{3}$$

True Negative: classifying normal data as normal data. This is referred to as specificity.

$$TruenegativeRate = \frac{\#Correct_Normal}{\#Correct} \tag{4}$$

Accuracy and *Precision* are the two other measures commonly used as performance metrics

$$Accuracy = \frac{\#Correct_Classifications}{\#all_instances} \tag{5}$$

$$Precision = \frac{\#Correct_Intrusions}{\#Instances_classified_as_Intrusions} \tag{6}$$

5 Results

We performed 10 fold cross validation and run each algorithm 10 times for each fold. PSO/ACO2 algorithm had a 10² particles i.e. 100 particles, and this algorithm ran for a maximum of 200 iterations per every rule discovered. We ran these experiments on every sample once by taking the fitness function as "Precision", and again by taking the fitness function as "Sens*Spec". We have also checked these results by changing the continuous optimizer from "PSO" to "DE".

Differential Evolution (DE) probably doesn't work well with a mix of nominal and continuous attributes.

LibSVM is a tool for Support Vector Machines, it helps users to easily apply SVM for their applications. We classified our data using C-Support Vector Classification. The results of PSO/ACO2 algorithm on the three sample sets is compared with LibSVM are given in the table below. We ran PSO/ACO2 on all the three samples first taking the fitness function as Precision, and then by changing the fitness function to "Sensitivity*Specificity", both outputs are also given in the table. At the end in Table 2, we have given the rules set, developed by PSO/ACO2 when run on 20%_Sample, at Fold 9.

Table 1. Accuracy of the sample datasets while using PSO/ACO2 & LibSVM
 Note: **Pr** - Precision; **S*P** - Sens * Spec;

DataSet	PSO/ACO2	PSO/ACO2	LibSVM	DE	DE
	(Pr)	(S*P)		(Pr)	(S*P)
10%_Sample	98.8±0.6	99.3±0.33	98.69	99.1±0.54	99.25±0.46
20%_Sample	99.65±0.23	98.0±1.26	93.45	98.65±0.92	97.3±1.1
25%_Sample	99.75±0.19	97.17±0.84	95.28	99.18±0.9	95.93±1.82

Table 2. Rule Set developed by PSO/ACO2 on 20% sample at Fold 9

Rule 1 : IF 'same_srv_rate: continuous' <= 0.368282067052106 'dst_host_srv_error_rate: continuous' >= 0.26650779032544014 THEN dos Quality: 1.0 (100,0)
Rule 2 : IF 'dst_host_count: continuous' >= 48.3546222578456 'dst_host_rerror_rate: continuous' >= 0.0365026802187019 THEN dos Quality: 1.0 (41,0)
Rule 3 : IF 'count: continuous' >= 25.08809296825722 THEN dos Quality: 0.99 (18,0)
Rule 4 : IF 'service: symbolic' = eco.i 'dst_host_count: continuous' <= 4.040591067519088 THEN probe Quality: 0.99 (16,0)
Rule 5 : IF 'is_guest_login: symbolic' >= 0.4909124927395499 'dst_host_count: continuous' >= 191.80817311303713 THEN r2l Quality: 0.99 (15,0)
Rule 6 : IF 'dst_host_count: continuous' <= 253.06229774043365 'dst_host_same_srv_rate: continuous' >= 0.4896632989835674 'dst_host_srv_diff_host_rate: continuous' <= 0.09604194645830486 'dst_host_error_rate: continuous' <= 0.27186632414196266 THEN normal Quality: 0.99 (10,0)
Rule 7 : IF 'service: symbolic' = ftp_data THEN r2l Quality: 0.98 (5,0)
Rule 8 : IF 'service: symbolic' = private THEN dos Quality: 0.96 (2,0)
Rule 9 : IF 'count: continuous' <= 2.1697958258857852 'dst_host_count: continuous' >= 40.3431922437951 'dst_host_same_srv_rate: continuous' >= 0.40088245887596413 THEN normal Quality: 0.95 (2,0)
Rule 10 : IF 'service: symbolic' = imap4 THEN r2l Quality: 0.91 (1,0)
Rule 11 : IF 'flag: symbolic' = s0 'dst_host_srv_count: continuous' <= 56.65140970706417 THEN dos Quality: 0.86 (1,0)
Rule 12 : IF 'service: symbolic' = http THEN normal Quality: 0.8 (0,0)
Rule 13 : IF 'dst_host_count: continuous' <= 238.19319375448072 THEN normal Quality: 0.67 (0,0)
Rule 14 : IF 'service: symbolic' = ecr.i THEN dos Quality: 0.23 (1,0)
Rule 15 : IF THEN r2l Quality: 0.05 (0,0)

6 Conclusions and Future Work

Intrusion detection systems are complex systems which include various modules, and we need several theories and techniques to efficiently fuse these models and get satisfying results.

In the present paper we are looking mainly for implementing Swarm Intelligence algorithms as intrusion detection engine in IHDAIDS, this proved as efficient in classifying normal data instance and attacking instances. From the

chosen algorithms, PSO/ACO2 is the best compared with Differential Evolution and LibSVM algorithms when the attack size increases.

We also proposed that an intrusion detection system should be intelligent, hybrid, adaptive and active to efficiently identify intrusions and deter the adversaries / hackers.

References

1. Cavusoglu, H., Mishra, B., Raghunathan, S.: The Value of Intrusion Detection Systems in Information Technology Security Architecture. *Information Systems Research* 16(1), 28–46 (2005)
2. Zhou, T.-J., Li, Y., Li, J.: Research on Intrusion Detection of SVM Based on PSO. In: *Proceeding of the Eighth International Conference on Machine Learning and Cybernetics*, Baoding (2009)
3. Holden, N., Freitas, A.A.: A hybrid particle swarm/ant colony algorithm for the classification of hierarchical biological data. In: *Proceeding of 2005 IEEE Swarm Intelligence Symposium (SIS 2005)*, pp. 100–107. IEEE (2005)
4. Holden, N., Freitas, A.A.: Hierarchical Classification of G-Protein-Coupled Receptors with a PSO/ACO Algorithm. In: *Proceedings of IEEE Smarm Intelligence Symposium (SIS 2006)*, pp. 77–84 (2006)
5. KDD Data Sets (1999),
<http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>
6. Spaord, E., Zamboni, D.: Data Collection Mechanisms for Intrusion Detection Systems, CERIAS Technical Report (2000)
7. Bashah, N., Shanmugam, I.B., Ahmed, A.M.: Hybrid Intelligent Intrusion Detection System. *World Academy of Science, Engineering and Technology* (2005)
8. Srinoy, S.: Intrusion Detection Model Based On Particle Swarm Optimization and Support Vector Machines. In: *Proceedings of the IEEE Symposium on Computational Intelligence in Security and Defence Applications, CISDA 2007* (2007)
9. Tavallae, M., Bagheri, E., Lu, W., Ghorbani, A.A.: A detailed Analysis of KDD CUP 99 Data set. In: *Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defence Applications (CISDA 2009)* (2009)
10. Peddabachigri, S., Abraham, A., Grosan, C., Thomas, J.: Modelling Intrusion Detection System using Hybrid Intelligent System. *Journal of Network and Computer Application* 30, 114–132 (2007)
11. Mukkamala, S., Janoski, G., Sung, A.: Intrusion Detection using Nerual Networks and Support Vector Machines. IEEE (2002)
12. Price, K., Stron, R., Lampinen, J.: *Dierential Evolution*. Springer, Heidelberg (2005)
13. Onwubolu, G.C., Devendra, D.: *Dierential Evolution: A handbook on Global Permutation Based Combinatorial Optimization*. Springer, Heidelberg (2009)

Genetically Optimized Supplementary Controller for SSSC to Damp Subsynchronous Oscillations

Sasmita Padhy and Sidhartha Panda

¹ Department of Electrical and Electronics Engineering,
National Institute of Science and Technology (NIST),
Berhampur, Odisha 761 008, India
panda_sidhartha@rediffmail.com

² Department of Electrical and Electronics Engineering,
Veer Surendra Sai University of Technology (VSSUT),
Burla, Odisha 768018, India
sasmita_padhy@hotmail.com

Abstract. A supplementary subsynchronous damping controller is proposed for the static synchronous series compensator (SSSC) device to damp out subsynchronous oscillations in power systems with series compensated transmission lines. The design problem of the proposed controller is formulated as an optimization problem, and real coded genetic algorithm is employed to search for the optimal controller parameters. It is shown that the controller is able to stabilize all unstable modes. The IEEE Second Benchmark Model is considered as the system under study. All the simulations are carried out in MATLAB/SIMULINK environment.

Keywords: Subsynchronous resonance, torsional oscillations, static synchronous series compensator, real coded genetic algorithm.

1 Introduction

Series capacitive compensation was introduced decades ago to cancel a portion of the reactive line impedance and thereby increase the transmittable power. Additionally it also improves the transient and steady state stability limits of power systems. However, capacitors may cause subsynchronous resonance (SSR) problems which result from the interaction between an electrical mode of the series compensated network and a mechanical shaft mode of a turbine-generator group [1]. This leads to turbine-generator shaft failure and electrical instability at oscillation frequencies lower than the normal system frequency. SSR is a resonant condition, with frequency below the fundamental frequency, which is related to an energy exchange between the electrical and the mechanical system, coupled through the generator. SSR can be divided in two main groups [2]: steady state SSR (induction generator effect: IGE, and torsional interaction) and transient torques, also known under the name of torque amplification (TA). IGE is considered as a theoretical condition that unlikely can occur in a series compensated power system, thus, it will not be considered in this paper. SSR due to TI and TA are dangerous conditions that can lead to shaft damage

[3] and therefore must be avoided. To avoid SSR in power systems, the use of FACTS devices such as the static synchronous compensator (STATCOM) [4], the static synchronous series compensator (SSSC) [5], [6] or the thyristor controlled series capacitor (TCSC) [7] and static var compensator (SVC) [8] have been proposed.

In this paper, a comprehensive assessment of the effects of SSSC-based damping controller for damping SSR has been carried out. First a simple lead-lag structure based supplementary controller for SSSC is proposed. Then, real coded genetic algorithm (RCGA) based optimal tuning algorithm is used to optimize the parameters of the SSSC based controller. The design objective is to reduce the torsional oscillation of second bench mark model, subjected to three phase fault. Simulation results are presented to show the effectiveness of the proposed approach.

2 System under Study

The IEEE SBM system-1 model [9] is considered as the SSR studying system which is shown in Fig. 1. The system consists of a single generator connected to an infinite bus via two transmission lines, one of which is series-compensated. The mechanical system is typically constituted by several masses representing different turbine stages interconnected by elastic shafts. When a torsional mode is excited, the masses perform small amplitude twisting movements relative to each other. The phase angle of the generator mass becomes modulated, causing a variation in the stator flux. Depending on the series-compensated network, substantial modulation of the stator current will result. In particular, if the frequency of this oscillating current is electrically close to the resonance frequency of the series compensated network, undamped currents will result. The flux in the generator and the stator current will create an electrical torque that will act on the generator mass. In the present study, the mechanical system is modeled by 3-masses: mass 1 = generator; mass 2 = low pressure turbine (LP); mass 3 = high pressure turbine (HP). The subsynchronous mode introduced by the compensation capacitor after a three-phase fault has been applied and cleared excites the oscillatory torsional modes of the multi-mass shaft and the torque amplification phenomenon.

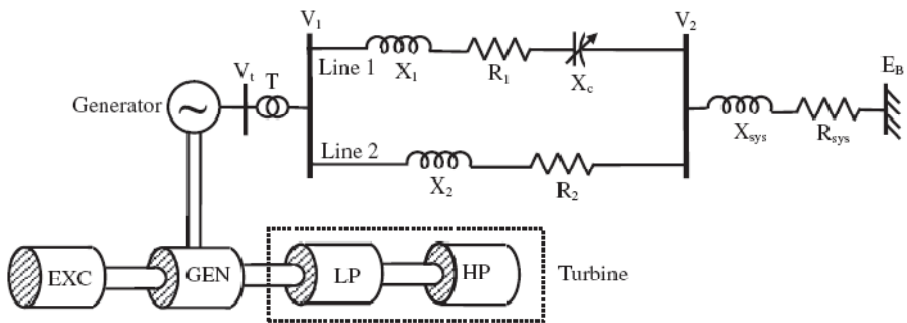


Fig. 1. IEEE Second bench mark (SBM) model

3 The Proposed Approach

3.1 Structure of the SSSC Based Supplementary Controller

The commonly used lead-lag structure shown in Fig. 2 is chosen in this study as a SSSC-based controller to modulate the SSSC injected voltage V_q . The lead-lag structure is preferred by the power system utilities because of the ease of on-line tuning and also lack of assurance of the stability by some adaptive or variable structure techniques. The structure consists of a gain block with gain K_S , a signal washout block and two-stage phase compensation block. The time delay introduced due to delay block depends on the type of input signal. For local input signals only the sensor time constants is considered and for remote signals both sensor time constant and the signal transmission delays are included. The signal washout block serves as a high-pass filter, with the time constant T_w , high enough to allow signals associated with oscillations in input signal to pass unchanged. From the viewpoint of the washout function, the value of T_w is not critical and may be in the range of 1 to 20 seconds [10]. The phase compensation blocks (time constants T_{1S} , T_{2S} and T_{3S} , T_{4S}) provide the appropriate phase-lead characteristics to compensate for the phase lag between input and the output signals. In Fig. 2, V_{qref} represents the reference injected voltage as desired by the steady state power flow control loop. The steady state power flow loop acts quite slowly in practice and hence, in the present study V_{qref} is assumed to be constant during the disturbance period. The desired value of compensation is obtained according to the change in the SSSC injected voltage ΔV_q which is added to V_{qref} .

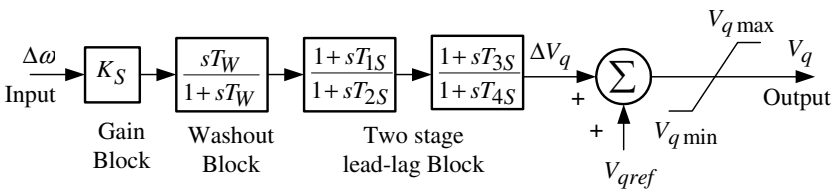


Fig. 2. Lead-lag structure of SSSC-based controller

3.2 Problem Formulation

In the lead-lag structured controllers, the washout time constants T_w is usually pre-specified [10]. A washout time constant $T_w = 10s$ is used in the present study. The controller gain K_S and the time constants T_{1S} , T_{2S} , T_{3S} and T_{4S} are to be determined. During steady state conditions ΔV_q and V_{qref} are constant. During dynamic conditions the series injected voltage V_q is modulated to damp system oscillations. The effective V_q in dynamic conditions is given by:

$$V_q = V_{qref} + \Delta V_q \tag{1}$$

It is worth mentioning that the SSSC-based controller is designed to minimize the torsional oscillations after a disturbance so as to reduce the damping due to sub

synchronous resonance. Minimization of these deviations could be chosen as the objective. In the present study, an integral time absolute error of the sum of the deviation errors signals corresponding to all modes of oscillations is taken as the objective function J as given below:

$$J = \int_{t=0}^{t=t_{sim}} |\Delta\omega_i| \cdot t \cdot dt \quad (2)$$

Where, $\Delta\omega$ is the deviation between two variables, t_{sim} is the time range of the simulation. For objective function calculation, the time-domain simulation of the power system model is carried out for the simulation period. It is aimed to minimize this objective function in order to improve the system response in terms of the settling time and overshoots. The problem constraints are the SSSC controller parameter bounds. Therefore, the design problem can be formulated as the following optimization problem:

$$\text{Minimize } J \quad (3)$$

Subject to

$$\begin{aligned} K_S^{\min} &\leq K_S \leq K_S^{\max} \\ T_{1S}^{\min} &\leq T_{1S} \leq T_{1S}^{\max} \\ T_{2S}^{\min} &\leq T_{2S} \leq T_{2S}^{\max} \\ T_{3S}^{\min} &\leq T_{3S} \leq T_{3S}^{\max} \\ T_{4S}^{\min} &\leq T_{4S} \leq T_{4S}^{\max} \end{aligned} \quad (4)$$

4 Overview of Real Coded Genetic Algorithm

Genetic algorithm (GA) has been used to solve difficult engineering problems that are complex and difficult to solve by conventional optimization methods. GA maintains and manipulates a population of solutions and implements a survival of the fittest strategy in their search for better solutions. The fittest individuals of any population tend to reproduce and survive to the next generation thus improving successive generations. The inferior individuals can also survive and reproduce. Implementation of GA requires the determination of six fundamental issues: chromosome representation, selection function, the genetic operators, initialization, termination and evaluation function. Brief descriptions about these issues are provided in the following sections [11].

4.1 Chromosome Representation

Chromosome representation scheme determines how the problem is structured in the GA and also determines the genetic operators that are used. Each individual or chromosome is made up of a sequence of genes. Various types of representations of an individual or chromosome are: binary digits, floating point numbers, integers, real values, matrices, etc. Generally natural representations are more efficient and produce better solutions. Real-coded representation is more efficient in terms of CPU time and offers higher precision with more consistent results.

4.2 Selection Function

To produce successive generations, selection of individuals plays a very significant role in a genetic algorithm. The selection function determines which of the individuals will survive and move on to the next generation. A probabilistic selection is performed based upon the individual's fitness such that the superior individuals have more chance of being selected. There are several schemes for the selection process: roulette wheel selection and its extensions, scaling techniques, tournament, normal geometric, elitist models and ranking methods.

4.3 Genetic Operator

The basic search mechanism of the GA is provided by the genetic operators. There are two basic types of operators: crossover and mutation. These operators are used to produce new solutions based on existing solutions in the population. Crossover takes two individuals to be parents and produces two new individuals while mutation alters one individual to produce a single new solution. The following genetic operators are usually employed: simple crossover, arithmetic crossover and heuristic crossover as crossover operator and uniform mutation, non-uniform mutation, multi-non-uniform mutation, boundary mutation as mutation operator. Arithmetic crossover and non-uniform mutation are employed in the present study as genetic operators. Crossover generates a random number r from a uniform distribution from 1 to m and creates two new individuals.

4.4 Initialization, Evaluation Function and Stopping Criteria

An initial population is needed to start the genetic algorithm procedure. The initial population can be randomly generated or can be taken from other methods. Evaluation functions or objective functions of many forms can be used in a GA so that the function can map the population into a partially ordered set. The GA moves from generation to generation until a stopping criterion is met. The stopping criterion could be maximum number of generations, population convergence criteria, lack of improvement in the best solution over a specified number of generations or target value for the objective function.

5 Results and Discussions

The SimPowerSystems (SPS) toolbox is used for all simulations and SSSC-based damping controller design [12]. SPS is a MATLAB-based modern design tool that allows scientists and engineers to rapidly and easily build models to simulate power systems using Simulink environment. In order to optimally tune the parameters of the SSSC-based damping controller, as well as to assess its performance, the model of example power system shown in Fig. 1 is developed using SPS blockset.

5.1 Application of RCGA

For the purpose of optimization of equation (3), RCGA is employed. For the implementation of GA normal geometric selection is employed which is a ranking selection function based on the normalized geometric distribution. Arithmetic crossover takes two parents and performs an interpolation along the line formed by the two parents. Non uniform mutation changes one of the parameters of the parent based on a non-uniform probability distribution. This Gaussian distribution starts wide, and narrows to a point distribution as the current generation approaches the maximum generation.

The objective function is evaluated for each individual by simulating the example power system, considering a severe disturbance. For objective function calculation, a 3-phase short-circuit fault is considered. An initial population is needed to start the genetic algorithm procedure. One more important point that affects the optimal solution more or less is the range for unknowns. For the very first execution of the program, more wide solution space can be given and after getting the solution one can shorten the solution space nearer to the values obtained in the previous iteration. Optimization was performed with the total number of generations set to 30. The optimization processes is run 10 times for both the control signals and best among the 10 runs are as follows:

$$K_S = 183.5414, T_{1S} = 0.26, T_{2S} = 0.1107, T_{3S} = 1.0714, T_{4S} = 0.1214$$

5.2 Simulation Results

The results from the simulation of the nonlinear system under study including the proposed SSSC based supplementary subsynchronous damping controller (SSDC) are shown in this section. All the simulations are carried out in MATLAB–SIMULINK environment for 55% compensation ratio of line-1. A self clearing 3-phase fault of 169 ms duration is applied near the sending end at $t = 0$ sec. Variation of generator speed deviation, LP-Generator speed deviation and HP-Generator speed deviation are shown in Figs. 3-5. In all the Figures, the response without the SSSC based supplementary subsynchronous damping controller (SSDC) is shown with dotted lines and the response with proposed SSSC based SSDC is shown with solid lines with. Please note that, in both the cases SSSC is present in the system. It is clear that the system is highly oscillatory for the above contingency without SSDC in all cases and all the subsynchronous oscillations are effectively damped out using the proposed RCGA optimized SSDC.

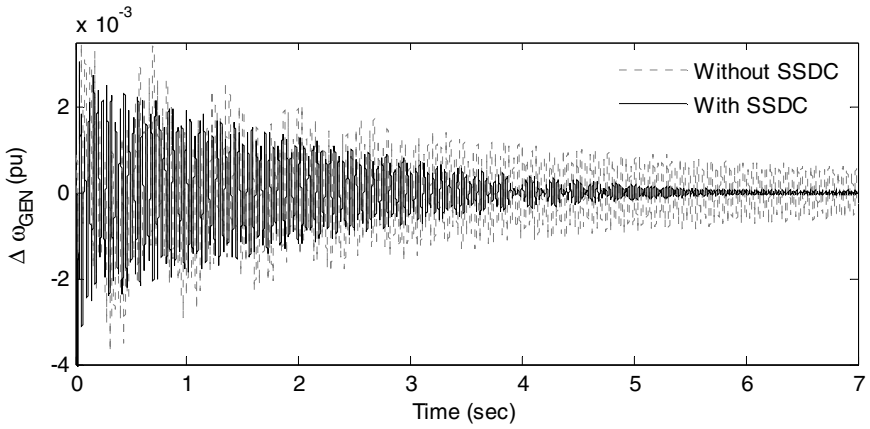


Fig. 3. Generator speed deviation with and without SSDC

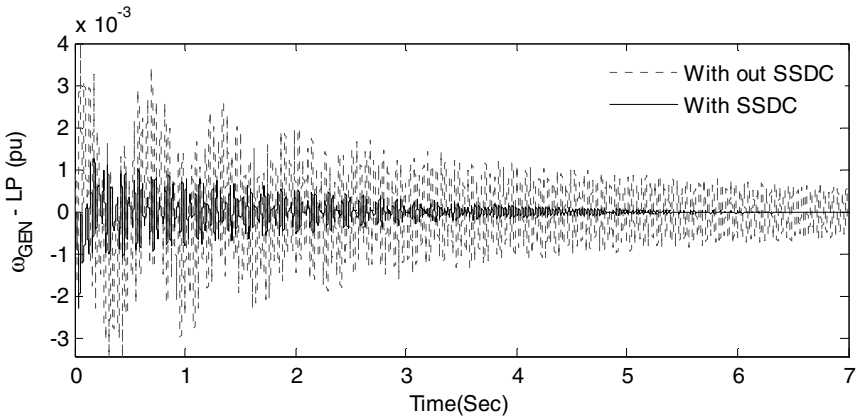


Fig. 4. Generator-LP speed deviation with and without SSDC

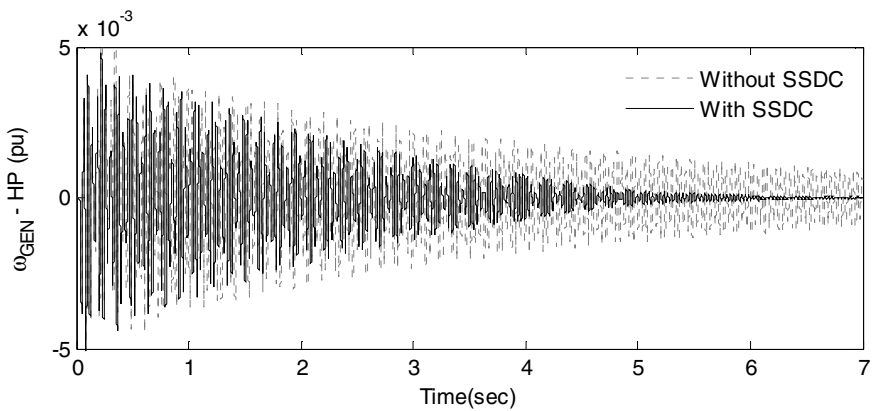


Fig. 5. Generator-HP speed deviation with and without SSDC

6 Conclusion

In this study, damping of subsynchronous resonance and low frequency power oscillation in a series compensated transmission line by a static synchronous series compensator (SSSC)-based supplementary subsynchronous damping controller is thoroughly investigated. The design problem of the proposed SSDC controller is formulated as an optimization problem with a simulation based objective function consisting of all oscillating modes. Then, real coded genetic algorithm (RCGA) is employed to search for the optimal controller parameters. The performance of the proposed SSDC controller is evaluated under a severe 3-phase fault disturbance. The simulation results show that the proposed SSDC controller is able to stabilize all unstable modes.

References

1. IEEE SSR Working Group, Terms, Definitions and Symbols for Subsynchronous Oscillations. *IEEE Trans. Power Appl. Syst.* 6, 1326–1334 (1985)
2. Anderson, P., Agrawal, B., Ness, J.V.: *Subsynchronous Resonance in Power Systems*. IEEE Press, New York (1989)
3. Padiyar, K.R.: *Analysis of Subsynchronous Resonance in Power systems*. Springer, Heidelberg (1998)
4. Padiyar, K.R., Prabhu, N.: Design and Performance Evaluation of Damping Controller with STATCOM. *IEEE Trans. Power Del.* 21, 1398–1405 (2006)
5. Kumar, L.S., Ghosh, A.: Modeling and Control Design of a Static Synchronous Series Compensator. *IEEE Trans. Power Del.* 14, 1448–1453 (1999)
6. Ooi, B.T., Dai, S.Z.: Series-type Solid-State Static VAR compensator. *IEEE Trans. Power Electron* 8, 164–169 (1993)
7. Perkins, B.K., Iravani, M.R.: Dynamic Modeling of a TCSC with Application to SSR analysis. *IEEE Trans. Power Syst.* 12, 1619–1625 (1997)
8. Padiyar, K.R., Varma, R.: Damping Torque Analysis of Static VAR System Controllers. *IEEE Trans. Power Syst.* 6, 458–465 (1991)
9. IEEE Subsynchronous Resonance Working Group: Second Benchmark Model for Computer Simulation of Subsynchronous Resonance. *IEEE Trans. Power Apparatus Syst.* 5, 1057–1066 (1985)
10. Kundur, P.: *Power System Stability and Control*. McGraw-Hill (1994)
11. Panda, S., et al.: Design and Analysis of SSSC-based Supplementary Damping Controller. *Simulation Modelling Practice and Theory* 18, 1199–1213 (2010)
12. SimPowerSystems 5.2.1: <http://www.mathworks.com/products/simpower/>

Index Page Synthesis Using Genetic Algorithm

Ashok Kumar Panda¹, Satchidananda Dehuri², and Isha Padhy¹

¹ Dept. of Computer Sc & Engg, MITS, Rayaada
akpanda7@yahoo.co.in, satchi.lapa@gmail.com

² Dept. of Comm. & IT, FM University, Balesore
ishapadhy06@gmail.com

Abstract. Evolutionary learning algorithms, such as Genetic Algorithms (GA) have been applied to information retrieval (IR) since 1980s. K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k-clusters) fixed a priori. We recommend on the design of fitness functions for genetic-based information retrieval experiments. This paper focuses on the problem of index page synthesis where an index page consists of a set of links that cover a particular topic. Its basic approach is to analyze web access logs, find groups of pages that often occur together in user visits and convert them into index pages. Grouping of pages can efficiently be done by clustering method. In this paper we have used the K-means algorithm for the purpose. Also considered is a fitness function that will find out the fitness value of a document & thus clusters are formed. The results indicate that, the design of fitness functions is instrumental in performance improvement.

Keywords: genetic algorithm, information retrieval, k-means algorithm, index page, fitness function.

1 Introduction

Document clustering has become an increasingly important technique for enhancing search engine results, web crawling, unsupervised document organization, and information retrieval or filtering. Since clustering is a known NP-hard problem (Garey and Johnson, 1979), most approaches use the alternative optimization schemes in order to find a local optimum solution of their criterion function[5]. Many stochastic optimization schemes that aim at global optimum have been reported, among which are the genetic algorithms (GAs). GAs are search procedures that use the mechanics of evolution and natural genetics.

Because of the intrinsic parallel search mechanism and powerful global exploration capability in a high-dimensional space, both GA and GP have been used to solve a wide range of hard optimization problems that often times have no best known solutions. The application areas cover a wide range of IR topics such as document indexing; query induction, representation, and optimization; document clustering; and document matching and ranking[3]. Suppose two ranking functions both retrieve 5

relevant documents in the top 10 results. Their relevance information (1 being relevant and 0 being non-relevant) in the ranking list are shown as follows:

Rank list 1 : 1 0 1 1 0 0 1 0 0 1

Rank list 2 : 1 1 0 1 1 1 0 0 0 0

If ranking order is ignored, the performance of the these two rank lists is the same — both returning 5 relevant documents. However, we should prefer the second rank list over the first because the second rank list presents relevant documents sooner (i.e., has higher precision). Our aim in this work has been to propose a clustering methodology which will be conceptually simple like the k-means algorithm. It should not suffer from the limitation of the K-means algorithm. We need to test whether this design will lead to good search performance. Here we try to find some high quality clusters that may be required for a particular query.

2 Problem Statement

Web mining can be broadly defined as the discovery and analysis of useful information from the WWW. Depending on the location of the source the type of collected data differs. We also have to focus on handling context sensitive and imprecise queries, and consider the need for personalization and learning.

2.1 Difficulties in Information Retrieval

- The aim of an IR system is to estimate the relevance of documents to users' information needs, expressed by means of queries[4]. Query processing in search engines is simple blind keyword matching. This does not take into account the context and relevance of queries with respect to documents.
- Page ranks are important since it is difficult to scan through the entire list of documents returned by the search engine in response to his/her query.

GAs, a biologically inspired technology, is randomized search and optimization techniques guided by the principles of evolution and natural genetics. They are efficient, adaptive and robust search processes, producing near optimal solutions. Here we have applied GA technique for analysis of web logs and to find index pages so that information retrieval can be done in effective way [2]..

3 Genetic Algorithm Approach

Relevancy is one of the major factor that has to be considered for getting relevant documents. We have attempted to group related relevant documents into clusters so that searching becomes an easy process.

3.1 Calculation of Relevancy

The chromosomes are strings of 0 and 1. 0 and 1 value is considered depending on the number of times a word is occurring in the document. The relevancy is considered by

calculating the number of 1's in the string. If the number of 1's are greater than some considered value then it is considered as relevant and the document is ranked accordingly. Length of the chromosome is assumed as 50.

Ex. The searched words are “discussion of genetic algorithm”
 Document 1: 1001
 Document 2:0001
 Document 3: 1111
 Document 1 and 3 are relevant.

3.2 The Problem Analysis- Documents Clustering

The searching capability of GAs has been used in this article for the purpose of appropriately determining a fixed number of K cluster centres in R^N

$$P(di) = \sum_{j=1}^i r(d_j) / i \tag{1}$$

Here we have considered the parameters as,

$r(d_i) = 0$, if document is not relevant
 1, if document is relevant
 $i = 1, 2, \dots, n$

The basic steps of GAs, which is followed here is,

String representation: Each string is a sequence of 0s and 1s representing the k-cluster centres. The sequence is the ranking sequence of D number of documents.

Fitness computation: The fitness computation process involves two steps: In the first phase the K- clusters are formed according to the ranking sequence encoded in the chromosomes under consideration. In the second phase we calculate the cluster centre by finding out the fitness value of the document.

For example, given a 1 0 1 1 ranking sequence, substituting i with 1, 2, 3, 4 in the equation we get 1, 0.5, 0.67, 0.75. It is easy to see that the fourth ranked document has higher value than the third ranked relevant document [6],[7].

Selection: On arrival of a new document , Its fitness value is calculated and the cluster is selected by applying K-means algorithm.

- k: number of clusters
- n_j : number of points in j^{th} cluster
- x_i^j : i^{th} point in j^{th} cluster
- $\min \sum_{j=1}^k \sum_{i=1}^{n_j} \|x_i^j - c_j\|^2$

The similarity between two documents must be measured in some way if a clustering algorithm is to be used. Here we have used K-means algorithm for defining similarity between two documents [10],[11].

Crossover : Crossover is a probabilistic process that exchanges information between two parent chromosomes for generating two child chromosomes. Here single point crossover with a fixed crossover probability of μ_c is used. For chromosomes of length l , a random integer, called the crossover point, is generated in the range $[1, l-1]$. The portions of the chromosomes lying to the right of the crossover point are exchanged to produce two offspring.

Mutation : Each chromosome undergoes mutation with a fixed probability μ_m , for binary representation of chromosomes, a bit position is mutated by flipping its value.

Termination criterion: Here the process of fitness computation, selection, crossover, and mutation are executed for a maximum number of iterations. The best string, seen up to the last generation, provides the solution to the clustering problem.

4 Literature Review

The document retrieval system has been studied through various papers for relevance information using fitness function in genetic algorithm. Robertson, A. M. et.al.[3] in 1994 evaluated the efficacy of a genetic algorithm with various order-based fitness functions for relevance feedback. Through another paper[4] in 1996 they described the development of a genetic algorithm for the assignment of weights to query terms in a ranked-output document retrieval system. The GA involves a fitness function that is based on full relevance information. In a paper in the same year [5] they proposed a novel nonlinear ranking function representation scheme and compare this new design to the well-known Vector Space model. During 90s the work on document relevancy through genetic algorithm has been found enormous success and the search engine so framed thereby the access pattern got to a new dimension. Smith M.P. et.al. [8] in the paper in 1997 discussed on Boolean query for an information need, given a small corpus of test documents, and then to use that query on the full collection to retrieve yet more relevant documents. Also Vrajitoru, V. in 1998[9] proposed that GA is to help an IR system to find, in a huge documents text collection, a good reply to a query expressed by the user.

The early period of 2000 show more work in this direction where the methods of clustering , document filtering etc was focused more for finding better relevancy to the documents searched. Mitchell, T. M. et.al. [1] in 2000 proposed a novel ranking function discovery framework based on Genetic Programming and shown how this helps automate the ranking function design/discovery process. In Proceedings of the 33rd -HICSS, Hawaii, in 2000[2] it is found that that possibility of applying GA's to adapt various matching functions lead to a better retrieval performance of documents. In the year 2004 , Areibi, S. and Z. Yang. [12] proposed a hybrid genetic algorithm for k -medoids clustering. A novel heuristic operator is designed and integrated with the genetic algorithm to fine-tune the search. Bandyopadhyay, S. and U. Maulik. in 2002 [13] presented a genetic algorithm for selecting centers to seed the popular k -means method for clustering. Tapas Kanungo et.al. in 2002[14] applied k -means clustering algorithm and presented a simple & efficient filtering algorithm. And in the year 2004 in a paper [15] Weiguo Fan et.al. have contrasted different fitness function designs on GP-based learning using a very large Web corpus.

5 The Algorithm

The algorithm proposed for the generation of relevant documents using k-means clustering is as follows:

```

Begin
  For each of the document, repeat
    Begin
      Calculate relevancy  $r(d)$  of the document;
    End for
  let the clusters taken initially be  $k$  ;
  Initialize interval to the greatest interval;
  Initialize TRel to total number of relevant documents
  While  $di > 0$  repeat
    Begin
      
$$P(di) = \sum_{j=1}^i r(d_j) / i$$

      Store  $P(di)$ 
      Initialize flag=0;
      Initialize  $l$  to  $k$ 
      While  $l > 0$  then
        If  $P(di)$  closest to  $cl$  then
          Add  $di$  to  $l$ 
          Calculate  $cl = \text{mean}(cl, P(di))$ 
        Flag=1
      End if
      If flag=0 then
        Initialize  $k+1$  as a new cluster and  $di$  as the centroid of the
        new cluster.
      End if
    End while
  End.

```

Every retrieved document under goes for a function $r(d)$ that gives a relevancy factor that will help to cluster the document in a more simplified manner and on which our algorithm is based on. Initializing the cluster size to a particular numeric value ' k '. then the Euclidean distance for each cluster, **Trel** is initialized with the value of total no of document available. The loop continues till the last document left out. Then calculates the precision value based on the relevancy factor that we have calculated in step 2 to 5. The values are stored to an array. Flag is initialized to 0 as an indicator that a document has been put in a cluster or not. If Flag is 0 then the document has not been added to any cluster as it varies with most attributes and is given by the interval value. l' is initialized to the no of currently created clusters. i.e k no of clusters. The while loop continues for every cluster with a centroid c_l . Then calculates the best fit l^{th} cluster for the current i^{th} document and it is added. The flag is set to 1 as it has been added to a particular cluster. If the flag is found to be 0 that indicates the new document is very different from the previous clustered documents and is set to be a new cluster. Expected search results with 10 documents and the document ids with 1,2,3....10. are shown as below.

6 Expected Results

Suppose we have a search results with 10 documents and the document ids are 1,2,3....10., And if the flag is set to be 0 indicating a new document, is very different from the previous clustered documents and is set to be a new cluster . The results as found from the Table 1 below shows less relevancy for higher ranked documents. The results shows a comparison of relevancy with and without experiments.

Table 1. Comparison of relevancy with & without the experiment

Document Number (di)	Rank	Relevancy without experiment	Relevancy with experiment, P(di)
1	1	0.9	1
2	2	0.8	0.5
3	3	0.7	0.66
4	4	0.6	0.75
5	5	0.5	0.8
6	6	0.4	0.66
7	7	0.3	0.57
8	8	0.2	0.62
9	9	0.1	0.55
10	10	0	0.6

The Figure 1 shown below depicts that higher ranked documents may be less relevant. The X- axis represents rank of the document and the Y-axis representing relevancy. Thus two series of lining got out of these documents show relevancy with & without the experiments, P(di). And the result is that higher is the rank, less is the relevancy of the document searched.

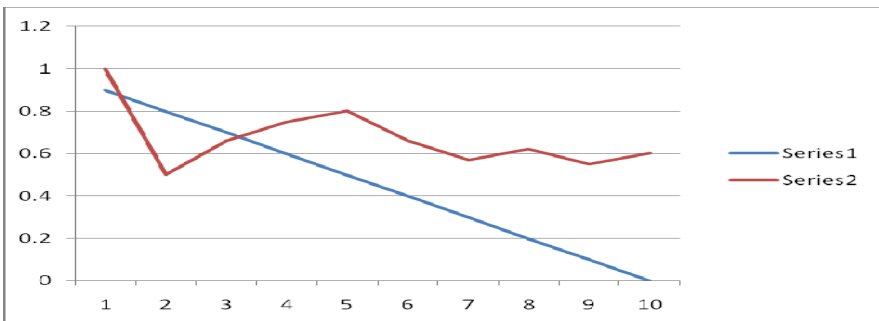


Fig. 1. Relevancy with and without experiments (P(di))

Table 2 below Calculates the best fit 1th cluster for the current ith document and it is being added to it. The flag is set to 1 as it has been added to a particular cluster. If the flag is found to be 0 that indicates the new document is very different from the previous clustered documents and is set to be a new cluster as found in place of finding a searched document. It is found from the algorithm taking 10 different documents so searched that, the documents of similar or close relevancies are clustered in a same group and having different relevancies in other cluster. And so formed four different clusters.

Table 2. Position of documents in a particular cluster

Cluster1	Cluster2	Cluster3	Cluster4
1	4,5	3,6,8,10	2,7,9

The figure 2 below depicts that, consecutive retrieved documents may not be closely equivalent ranked document that may be less relevant. The X-axis represents the document number(di) & the Y-axis the cluster number.

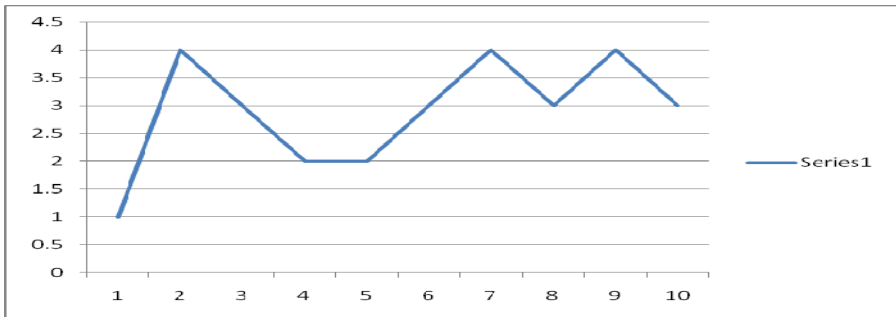


Fig. 2. Document -cluster representation

7 Conclusion

In this paper the problem of finding a globally optimal partition of a given set of documents into a specified number of clusters is considered. Here we tried to find some high quality clusters that may be required for a particular query. We extend our work to the implementation on search engines in near future and expect a good result. Genetic algorithm is a natural technique to find out the fittest chromosome that can survive, and clustering gives us the best result in finding out common featured chromosomes. So applying the concept in document clustering will help in getting the most appropriate document with respect to a searched word.

References

1. Mitchell, T.M., Pathak, P., Gordon, M., Fan, W.: Approach to an adaptive information retrieval agent. *Journal of the American Society for Information Science*. Machine Learning (2000)
2. Pathak, P., Gordon, M., Fan, W.: Effective information retrieval using genetic algorithms based matching function adaptation. In: *The Proceedings of 33rd Hawaii International Conference on System Science (HICSS)*, Hawaii, USA (2000)
3. Robertson, A.M., Willett, P.: Generation of equi-frequent groups of words using a genetic algorithm. *Journal of Documentation* (1994)
4. Robertson, A.M., Willett, P.: An upper bound to the performance of rank output searching-optimal weighting of query terms using a genetic algorithm. *Journal of Documentation* (1996)
5. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., Gatford, M.: Okapei at TREC-4. In: Harman, D.K. (ed.) *Proceedings of Fourth Text Retrieval Conference*, pp. 73–97. NIST Special Publication (1996) 500-236
6. Salton, G.: *Automatic Text Processing*. Addison-Wesley Publishing Co. (1989)
7. Reading, M.A., Salton, G., Buckley, C.: Term weighting approaches in automatic text retrieval. *Information Processing and Management* 24(5), 513–523 (1988)
8. Smith, M.P., Smith, M.: The use of genetic programming to build Boolean queries for text retrieval through relevance feedback. *Journal of Information Science* (1997)
9. Vrajitoru, V.: Crossover improvement for the genetic algorithm in information retrieval. *Information Processing and Management* (1998)
10. Agarwal, P., Sharir, M., Welzl, E.: The Discrete 2-Center Problem. In: *Proceedings of the 13th ACM Symposium on Computational Geometry*, pp. 147–155 (1997)
11. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. In: *Proceedings of the 20th VLDB Conference*, pp. 487–499 (1994)
12. Areibi, S., Yang, Z.: Effective Memetic Algorithms for VLSI Design Automation = Genetic Algorithms + Local Search + Multi-Level Clustering. *Evolutionary Computation*, 327–353 (2004)
13. Bandyopadhyay, S., Maulik, U.: An Evolutionary Technique Based on k-Means Algorithm for Optimal Clustering in RN. *Information Science*, 221–237 (2002)
14. Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y.: An effective k-means Clustering Algorithm: Analysis and Implementation. *IEEE Transaction on Pattern Analysis & Machine Intelligence* 24(7) (2002)
15. Fan, W., Fox, E.A., Pathak, P., Wu, H.: The Effects of Fitness function on Genetic Programming-Based Ranking Discovery For Web Search. *Journal of American Society for Information Science & Technology* 55(7), 628–636 (2004)

Web Shield: A Modified Firewall to Detect Malicious Request

Urjita Thakar^{1,*}, Omprakash Patel¹, and Lalit Purohit²

¹ Department of Computer Engineering, SGSITS, Indore
Uthakar@sgsits.ac.in, oppatel13@gmail.com

² Department of Information Technology, SGSITS, Indore
lpurohit@sgsits.ac.in

Abstract. In this paper, a novel system (Web Shield) for dynamic detection and filtering of objectionable web content has been presented. Web Shield captures the URL requested by the client and determines whether it leads to any objectionable content. It is also able to check all child URLs of the requested URL. It uses simple text matching technique to detect and filter undesired web pages. It uses knowledge base to keep record of objectionable URLs. The proposed system and the existing systems such as browser setting and proxy server based filtering were tested over about 500 web sites. It was observed that proposed system performed much better than existing methods giving about 90% accuracy. This system is very useful for various organizations enabling blocking and filtering of undesired and objectionable web content.

Keywords: Proxy Server, knowledge base, Content filtering, Image processing.

1 Introduction

The World Wide Web is growing very rapidly day by day. WWW is a huge information source and variety of services related to business, entertainment, news, educational, games etc are available on it. The openness of the web allows any user to access almost any type of information. However, all type of information is not appropriate for all users. Especially in an organization that may be software development organization or an educational organization, for better productive learning, the user should not be allowed to surf any content that could distract their attention. Also surfing of certain kind of bad content may cause losses to the organizations money, time, and their resources and also may harm user's mental strength. Thus it is challenge to prevent access to undesired information from the WWW.

Some companies and organizations are researching solution to this problem. The solutions have been focused on IP-based, URL-based and URL's content based filtering [1], image analysis and limited text filtering [2]. The classification of the Web sites is mostly manual but, as we know, the website is a highly dynamic

* Corresponding author.

information source. Not only many websites appear every day while others disappear, but also site content (including linkage information) is updated frequently. Thus, manual classification and filtering systems are largely impractical. Thus, it is a challenge to prevent access to undesired information from the WWW. The image and text analysis are more common for filtering porn web content [2], [3]. All objectionable content may not have images (e.g. proxy sites, share market news, some music and video downloading sites and games sites etc.) Thus these cannot be filtered on the basis of image analysis. Such dynamic character of the web and partial work in the field of all objectionable sites for any educational institutional an industrial organization calls for new techniques designed to classify and filter web sites and URLs automatically. Also, listing of either IP address or URLs name for filtering of objectionable content are grossly insufficient as the numbers of sites are increasing very rapidly. In this paper, we propose a dynamic objectionable content detection and filtering system (Web shield). The proposed system is able to detect and filter not only porn web content but also other sites that are deemed objectionable by an organization. Such sites include share market sites, news sites, music and video downloading sites, proxy sites, game downloading sites and many more.

2 Background

In past, some researchers have worked in the field of filtering and blocking of URLs. The most popular method for filtering is centralized content-based web filtering and blocking through use of proxy server [1], [3]. A well known proxy server for filtering and blocking centrally is a Squid proxy server. Squid proxy server can be configured as a network firewall in which the administrator defines some rules on the basis of IP address, domain name of URL and objectionable key words. These have only limited utility as they only perform static filtering. Forsyth et al proposed a method to detect porn image with the help of angle measurement [4]. Work related to more accurately detection of porn images has also been discuss in [5-7]. These methods were insufficient and only able for image analysis hence only able to perform porn web content filtering. Oardand et al proposed a frame work for text filtering [8]. Mohamed Hammami et al proposed a web filtering system based on face detection and text analysis using data- mining technique [2]. This method was suitable for porn web content but there are some more areas which comes under objectionable. Kazuhiro Iwahama et al proposed method for filtering System for Music Data [9]. Jos e de J. P erez et al proposed a method for information filtering system in web [10].

3 Proposed Web Shield Architecture

In this section, architecture of the Web shield has been presented. It has been designed as a proxy server with special feature to efficiently filter objectionable content. The system has following component as shown in figure 1.

1. **Conventional Proxy Server:** - It performs all functions as a conventional proxy server. It forward URLs to URL decision block.
2. **Database Decision Block (DDB):-** It matches URLs with the URLs already stored in the knowledge base. According to matching or miss- matching it takes appropriate decision for given URL that shall be fed to the HCG and main server or send an error message to client request.
3. **URL Queue:** - It is the queue of requested URLs. It is used for presenting URLs sequentially to HTML code generator as accessed by client.
4. **HTML Code Generator (HCG):** - HCG generates source code of web pages corresponding to the URL in text file. This source code of web page is used for text analysis.
5. **Parser and Database Query Processor:** - Parser is use to analyze source code of web page and find out objectionable word in it. In case any objectionable word is found then it is added to the knowledge base.
6. **Knowledge base:** - This knowledge base contains all the URLs that lead to a websites containing objectionable material.

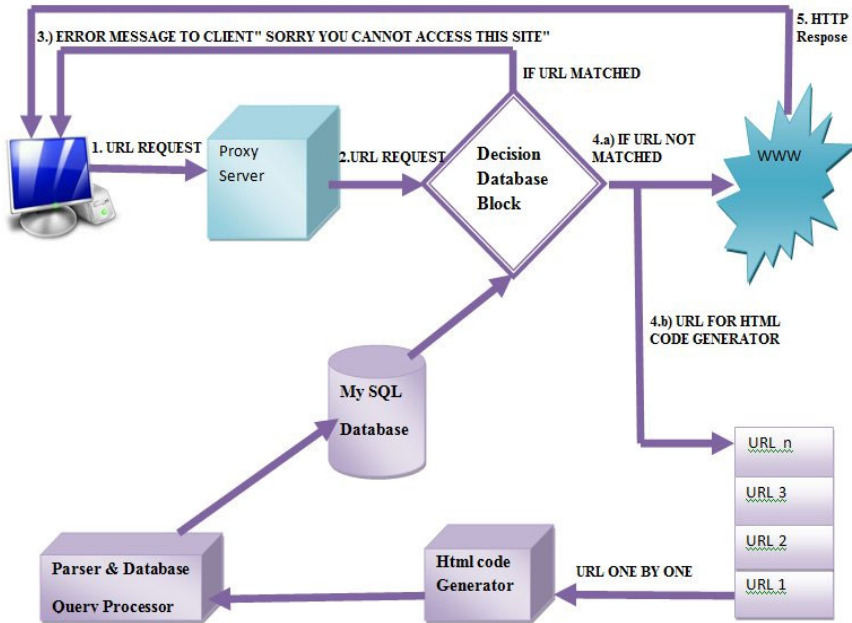


Fig. 1. Architecture of Web Shield

4 Working of Web Shield

Web shield detects and filters rapidly the Web pages with the help of its components. The client requests a URL from its browser. The requested URL is monitored by conventional proxy server. Proxy server forwards this requested URL

to DDB. DDB matches URL with URLs present in the knowledge base. If URL matches with a URL present in knowledge base then request denied message is send to client. If it does not match then URL (i.e. is not a URL leading to objectionable content) is forwarded to the Internet and also put in URL queue for its further processing. HTTP response of requested URL is displayed through browser to client. Mean while HCG generates source code of the requested web page in a text file. Web Shield has a special feature of checking all child URLs corresponding to requested URL. It is process of analyzing URL from top to bottom. To find out whether this source code contains any objectionable word the code is parsed. This also counts objectionable words frequency. If any objectionable words are found then the requested URL is inserted in to knowledge base.

In nut shell the tep followed are as given below:

WEB SHIELD (URL)

```
{
  1. Get URL from user through Proxy Server
  2. Match URL with database or knowledge base
  3. If found then send error message to user
     Else
  4. For each generate HTML code
  5. Get HTML code of that URL
  6. Parse the HTML code and find objectionable word
  7. If (objectionable word found)
     {
       Then make entry of that URL in database
     }
     Else
  8. Extract next requested URL from queue in HTML code generator
  9. Go to step no. 5
}
```

5 Evaluation of the Proposed System

In this section, evaluation of the existing and proposed methods has been presented. In the experiment 500 web sites were considered. The sites were from areas related to social networking, song downloading, proxy sites, game sites and some porn site etc. The existing methods included in the study that are related to browser setting, proxy server configuration.

5.1 Browser Setting

All browsers provide facility for listing URLs that are to be blocked. Figure 2 shows how the Internet explorer browser can be configured for this. Since according to the WWW survey, every day 20000 porn sites are coming up. It is not possible to list all the sites that are considered objectionable. Also, this configuration is to be done by the client at its end hence administrator has very little control on the setting. Hence it is not an effective mechanism.

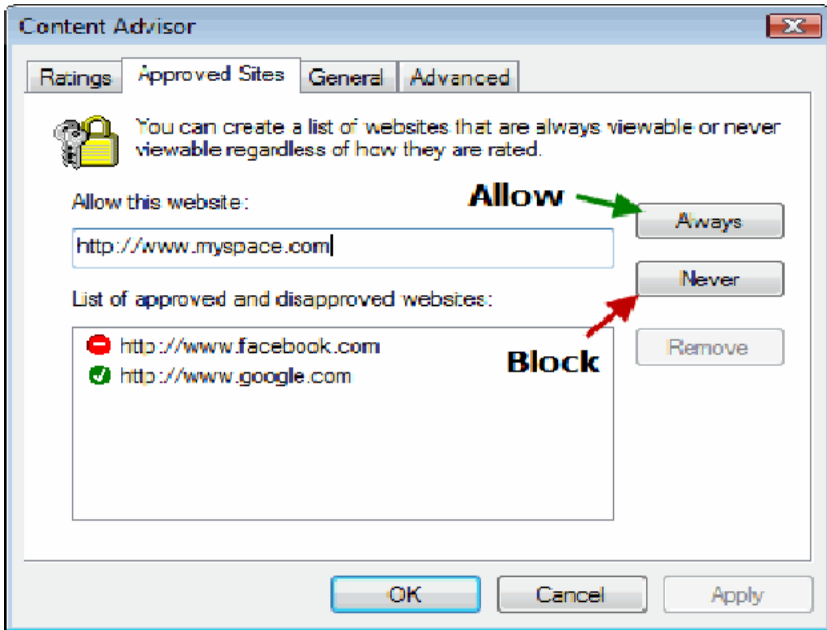


Fig. 2. Method to block URL through Browser

5.2 Proxy Server Configuration

In this experiment Squid proxy server has been used. It allows filtering and blocking of the URLs based on domain name, IP address and key words presented in the URLs. These are defined in the rules in access control list (ACLs). Samples of such access control rules are shown in figure 3, 4 and 5.

For IP address based blocking the, IP addresses of web sites are written in the ACL as shown in figure 3.

For domain name based blocking the domain name of web sites are written in the ACL as shown in figure 4.

For keywords based blocking the objectionable words that appear in URLs of the sites that are to be blocked are written in the ACL as shown in figure 5.

5.3 Proposed Web Shield

Dynamic nature of web is a major problem for filtering and blocking for undesirable URLs. Web shield uses a dynamic technique to detect and filter URL efficiently. Client can access any objectionable site only once and next time client will be unable to access that URL. It can be shown through following figure 6, 7. Let the user request a website www.iwin.com, in which there is no objectionable word in URL but its content is objectionable.

Initially this URL is not present in knowledge base. After accessing this site Web Shield process it to determine that the content of this sites are objectionable or not. Since it is gaming site which is objectionable for us. Therefore, Web Shield makes

an entry of this URL dynamically in the knowledge base. Web Shield has a special feature of checking all child URLs corresponding to requested URL, and saves, or discard children URL of the given URL depending on their content. The knowledge base after request to a objectionable URL is made is shown in figure 6. Figure 7 shows that on requesting same web page again, the client is denied access to the URL since it had objectionable content.

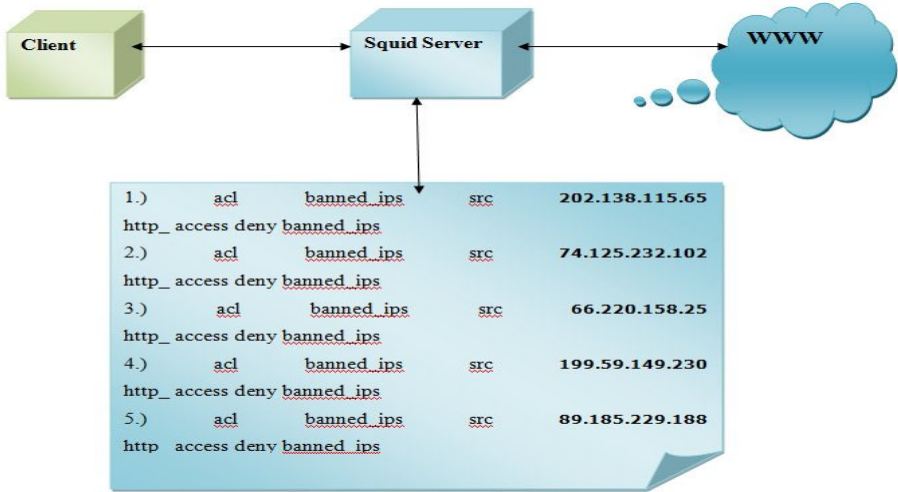


Fig. 3. Blocking and filtering through IP address of URL

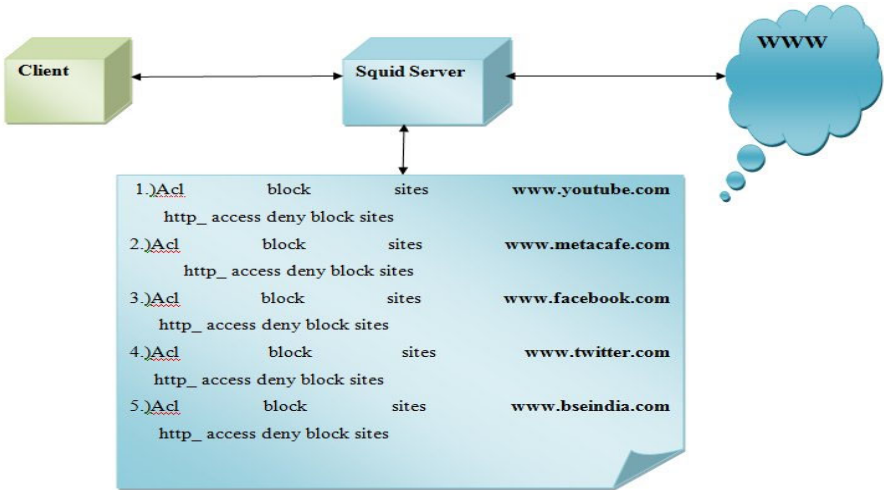


Fig. 4. Blocking and filtering through Domain name of URL

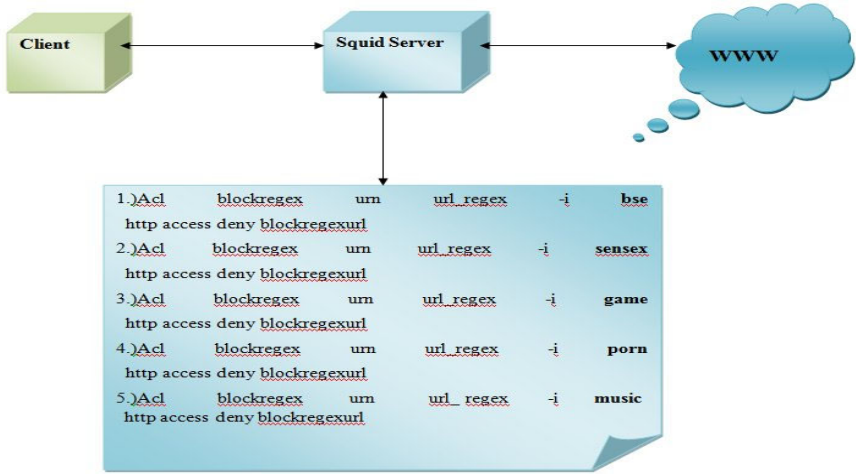


Fig. 5. Blocking and filtering of URL through keywords

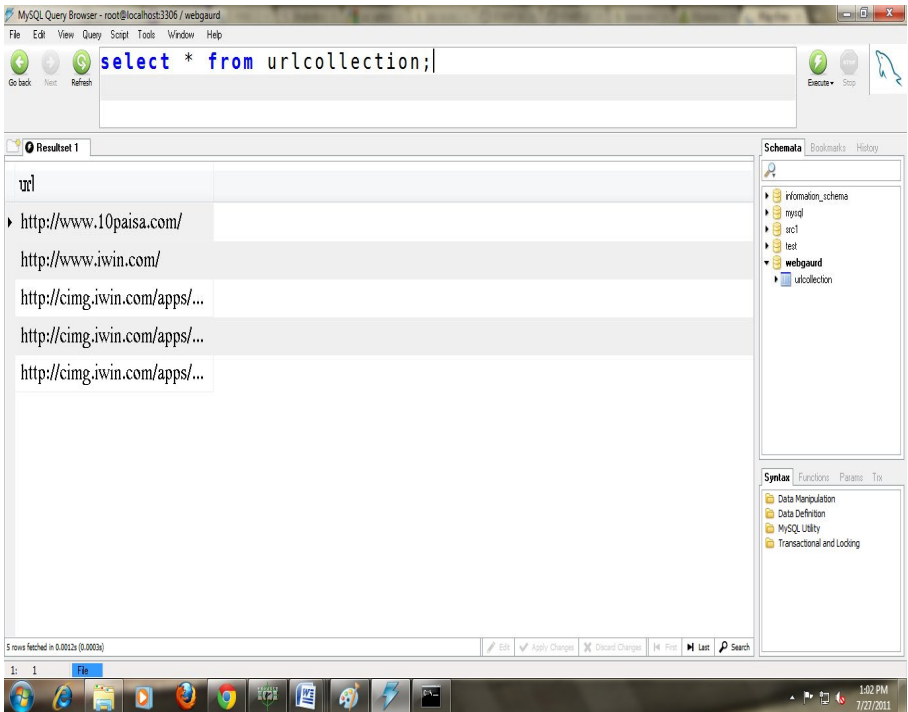


Fig. 6. Status of Knowledge Base after URL requested by client



Fig. 7. Result to client at browser, after requesting same site twice

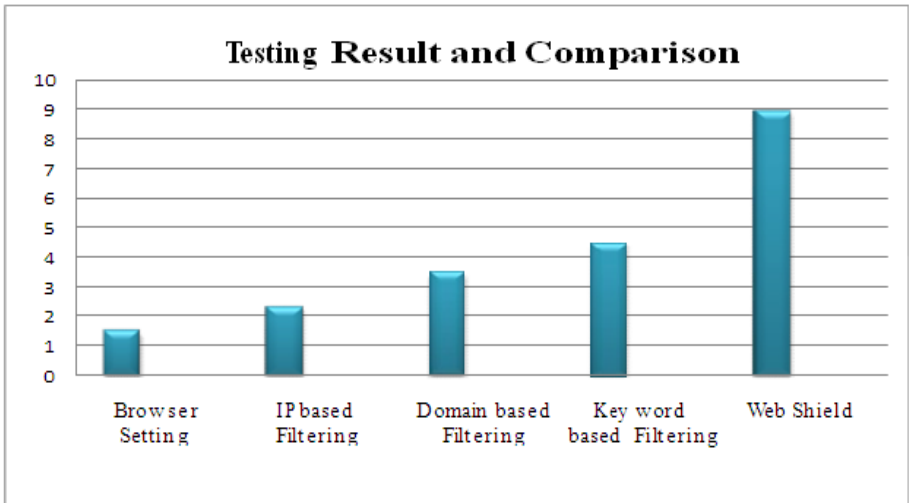


Fig. 8. Comparison Chart for different method

6 Discussion

The existing method of blocking undesirable URLs such as using the facilities in the browser setting and using conventional ACL's in proxy servers and the proposed Web Shield were evaluated in the previous section.

Browser setting, has limited capability to block and filter objectionable web content. These are performed at the user end and there is no control of administrator. It gives very poor result. It has capability to block and filter only up to 75 URLs out of 500 URLs if specified correctly in the browser.

The blocking and filtering through IP address, through proxy server is better than browser setting, as rule written in squid is applicable for all users of LAN. Nobody except administrator can modify the rule or the listed IP addresses. However it is not possible to write all IP addresses of various URLs, therefore this method can filter and block only 125 URLs out of 500 test objectionable URLs.

The domain name based blocking in Squid proxy server is better than IP based blocking, as it is easy to list domain name as compared to IP address. Thus, it gives better performance than IP address based blocking. It blocks only 175 URLs out of 500 objectionable URLs. This method also suffers from the limitation of inability of listing all the URLs and domain as new websites are being added very rapidly.

Keyword based blocking and filtering gives better result than above methods. It blocks 225 URLs out of 500 objectionable URLs. However, as it can detect only those URLs in which objectionable words that are listed in the rule are present.

There are several websites whose URL does not contain any objectionable word but their content is objectionable. The sites www.sensex.com and www.10paisa.com have same content. However, to block both sites the keywords present in both the URLs should be listed.

Also the ACL rule is case sensitive for key words therefore it is permitted to access objectionable site by just changing case of alphabet of keyword in URL. e.g. let [sensex](http://www.sensex.com), word be blocked but it can be access by entering word [Sensex](http://www.Sensex.com) where s has been capitalized. Otherwise the site, whose URL does not contain the listed word will be accessible.

The proposed Web Shield is a dynamic method to detect and filter web contents. Web shield has some limitations, that it is unable to analyze images. Therefore, it is unable to block all images which do not have any text clue. On testing the 500 test websites the result for the various methods are shown graphically in figure 8.

URL blocking using browser setting gives 15% accuracy. The proxy server IP based method gives 25% accuracy. The proxy server domain name of URL based blocking gives 35% accuracy. The proxy server keywords based method gives 45% accuracy. It can be clearly seen that the proposed Web Shield gives much better performance of blocking and filtering 450 URLs out of 500 objectionable URLs. Thus it gives result up to 90%. This method is also able to block the child URLs.

7 Conclusion and Future Work

In this paper, a 'Web Shield' modified firewall has been presented that is more efficient, dynamic and useful than the existing methods for blocking and filtering objectionable web content. The method uses simple text matching technique to detect and filter web pages. It is easy to increase or decrease objectionable keywords range by just updating keywords in parser of Web Shield. It is more advantageous than other method as it also filters and blocks the child URLs presented in the requested URLs. It does not hamper working at the client's end as it works at a central place.

The limitation of earlier methods such as number of IP addresses, domain names, keywords and case sensitivity of key words has been overcome. This method can be improved by adding image analysis technique and using log information of the user.

References

- [1] Feng, S., Zhang, J., Zeng, B.: Design of the Visualized Assistant for the Management of Proxy Server. Office of R&D, Naval Univ. of Eng, Wuhan, China (2009)
- [2] Hammami, M., Tsishkou, D., Chen, L.: Adult Content Web Filtering and Face Detection Using Data-Mining Based Skin-Color Model. LIRIS, Ecole Centrale de Lyon, France (2004)
- [3] Ding, C., Chi, C.-H., Deng, J., Dong, C.-L.: Centralized Content-Based Web Filtering and Blocking: How Far Can It Go? In: International Conference on Industrial Technology. IEEE (1999)
- [4] Fleck, M.M., Forsyth, D.A., Bregler, C.: Finding Naked People. Computer Science Division, U.C. Berkeley, Berkeley, CA 94720 (1996)
- [5] Girgis, M.R., Mohmoud, T.M., Abd-El-Hafeez, T.: An Approach to Image Extraction and Accurate Skin Detection From Web Pages. World Academy of Science in 2007 (2007)
- [6] Yang, J., Fu, Z., Tan, T., Hu, W.: A Novel Approach to Detect Objectionable Images. In: Proc. IEEE ICPR 2004, vol. 4, pp. 479–482 (2004)
- [7] Zeng, W., Gao, W., Zhang, T., Liu, Y.: Image Guarder: An Intelligent detector for Objectionable Images. In: Asian Conf. On Computer Vision, pp. 198–203 (January 2004)
- [8] Oardand, W., Marchionini, G.: A conceptual frame work for text filtering process, Technical Report CS-TR-3643, University of Maryland, College Park (1996)
- [9] Iwahama, K., Hijikata, Y., Nishida, S.: Content- based Filtering System for Music Data, 1-3 Machikaneyama, Toyonaka Osaka 560-8531, JAPAN (2006)
- [10] de J. Pérez-Alcázar, J., Calderón-Benavides, M.L., González-Caro, C.N.: Towards an Information Filtering System in the Web Integrating Collaborative and Content Based Techniques. Universidad Autónoma de Bucaramanga, Colombia (2003)

Optimal Samples Selection from Gene Expression Microarray Data Using Relational Algebra and Clustering Technique

Soumen Kr. Pati¹ and Asit Kr. Das²

¹ Department of Information Technology,
St. Thomas' College of Engineering and Technology,
4, D.H. Road, Kolkata-23

² Department of Computer Science and Technology,
Bengal Engineering and Science University, Shibpur, Howrah-03
{Soumen_pati, asitdas72}@rediffmail.com

Abstract. Real data of natural and social sciences is often very high-dimensional. Dataset handling in high-dimensional spaces presents complicated problems, such as the degradation of data accessing, data manipulating as well as query processing performance. Dimensionality reduction efficiently tackles this problem and benefited us to visualize the intrinsic properties hidden in the dataset. The proposed method first generates decision attribute by computing the class label of each gene using clustering technique and subsequently computes the score of each sample of microarray cancerous gene data based on decision attribute using the division operation of relational algebra and select the samples with score below the average score as initial reduct. The reduced dataset is grouped into k clusters by k -means algorithm where, k is the set of values of decision attribute and matching factor of reduct is computed by considering the overlapping of clusters with the original classes of genes. Other samples are added iteratively one at a time based on their increasing score provided computed matching factor improved and thus final reduct known as optimal set of samples is obtained.

1 Introduction

Now-a-days, an increasing number of applications in different fields especially on the field of natural and social sciences produce massive volumes of very high dimensional data [1] under a variety of experimental conditions. In scientific databases like gene microarray dataset, it is common to encounter large sets of observations, represented by hundreds or even thousands of coordinates. The performance of data analysis such as clustering and classification degrades in such high dimensional spaces. Gene microarray high dimensional data provides the opportunity to measure the expression level of thousands of genes simultaneously and this kind of high-throughput data has a wide application in bioinformatics research. In DNA microarray data [2] analysis generally biologists measure the expression levels of genes in the tissue samples from patients, and find explanations about how the

genes of patients relate to the types of cancers they had. Many genes could strongly be correlated to a particular type of cancer, however, biologists prefer to focal point on a small subset of genes that dominates the outcomes before performing in-depth analysis and expensive experiments with a high dimensional dataset. Therefore, automated selection of the small subset of attributes is highly advantageous.

Gene expression microarrays offer a popular technique to monitor the correlated expression of thousands of genes under a variety of experimental circumstances. In spite of the enormous potential of this technique, there remain challenging problems associated with the achievement and analysis of microarray data that can have a reflective influence on the interpretation of the outcomes. DNA microarray technology has directed the focus of computational biology towards analytical data interpretation [3]. However, when examining microarray data, the size of the data sets and noise contained within the data sets compromises precise qualitative and quantitative analysis[4]. Conventionally, a dimensionality reduction technique [5-6] may be used to reduce the size of the dataset before further processing. Dimensionality reduction can also provide a low level visual representation of gene behavior across the samples. A standard objective of microarray data analysis is to better understand the gene-to-gene interactions that take place amongst the entire gene pool. However, applications of dimensionality reduction techniques on microarray data have been only partially successful. Dimensionality reduction of microarray data has yet to effectively tackle the problem of finding a low dimensional embedding that provides a precise visual representation of gene-to-gene interactions [14]. Therefore, a fast and effective algorithm is needed for efficient processing of datasets of both high dimensionality and cardinality. In the article, a novel dimension reduction technique has been proposed that can be broken down into following four steps:

(i) The gene expression dataset is clustered and validated using [7] to achieve an optimal set of clusters and assign same class label to all the genes within a cluster so that two genes of different clusters have different class label. Thus decision attribute is generated for the dataset. Let k is the set of values of decision attribute.

(ii) The dataset is standardized to Z-score using Transitional State Discrimination method and each sample is characterized by four discrete values. After discretizing the gene expression data, score of each sample with respect to generated decision attribute is computed using division operation of relational algebra [8]. Lower the score implies more important sample of the gene dataset and vice versa.

(iii) Initial reduct is formed by considering the samples with score below the average score. The reduced dataset is grouped into k clusters by k-means algorithm and matching factor of reduct is computed by considering the overlapping of clusters with the original classes of genes. Other samples are added iteratively one at a time based on their increasing score provided computed matching factor improved and thus final reduct known as optimal set of samples is obtained.

The rest of the paper is organized as follows: Section 2 describes the proposed dimension reduction methodology to select only the relevant samples. The experimental results and performance of the proposed method for a variety of benchmark gene expression datasets is evaluated in Section 3. Finally, conclusions are drawn in Section 4.

2 Sample Selection and Dimension Reduction

A gene expression dataset is presented as a continuous gene expression matrix, where each row represents a gene and each column represents a sample that can be measured for each gene. Minimum subset of samples that preserve the homogeneity relation and classification power is known as reduct whereas other samples are irrelevant for classification and so removed from the dataset. Usually, there are several reducts of gene expression dataset finding to which is NP-hard problem [9]. Though the sample number is less compare to the number of genes still there exist some unimportant samples which increases time complexity while comparing characteristics of genes.

2.1 Relevance Analysis of Samples by Gene Clustering

As all samples are not relevant to characterize the genes, a relevance analysis of samples is necessary to select only the important samples. Here, a score is computed for each sample using division operation of relational algebra for which decision attribute is required. But the gene datasets are unlabeled data, as a result the datasets need to be transformed to labeled data. So, the dataset is clustered and validated using [7] to obtain optimal clusters of genes and labeled each gene in such a way that genes in same cluster have same class value and genes in different clusters have different class value. Thus, a decision attribute is generated which is used to compute score of each samples. Let the labeled gene expression dataset $DS = (U, C, D)$, where $C = \{C_1, C_2, \dots, C_n\}$ is the number of samples and D is the decision attribute with k distinct values generated using clustering algorithm. Score computation by division (\div) operation of relational algebra needs the discrete values of the samples. So, before score computation, the datasets are preprocessed by standardizing the samples to z-score using Transitional State Discrimination method (TSD). In TSD, discretization factor f_{ij} is computed for sample value $C_j \in C$ of gene $g_i \in U, i = 1, 2, \dots, m; j = 1, 2, \dots, n$, using (1).

$$f_{ij} = \text{round} \left(\frac{g_i[C_j] - \mu_i}{\delta_i} \right) \tag{1}$$

where, μ_i and δ_i are the mean and standard deviation of gene g_i and $g_i[C_j]$ is the value of sample C_j in gene g_i . Then the value $g_i[C_j]$ is discretized to one of ‘VL’ (very low), ‘L’ (low), ‘Z’ (zero), ‘H’ (high) and ‘VH’ (very high) depending on f_{ij} is ‘< -1’, ‘-1’, ‘0’, ‘1’ and ‘>1’ respectively.

Now, the relational algebra operation division (\div), defined in ‘Definition 1’, is used to compute the score of each sample C_i using score function $S(C_i)$ defined in equation (2).

$$S(C_i) = |\Pi_{C_i \cup D}(DS) \div \Pi_D(DS)| \tag{2}$$

where $i = 1, 2, \dots, n$. Minimum score of C_i implies that there is maximum number of genes having sample values similar to C_i , which can uniquely take the decisions. Thus, the sample with minimum score is of maximum importance and so lower score implies higher possibility of becoming a member of reduct.

Definition 1: Relational algebra operation division (\div) is a binary operation applied on two relations $R_1(P)$ and $R_2(Q)$ and produce another relation $R(P - Q)$ where $Q \subset P$ where P, Q are set of attributes of R_1, R_2 respectively. So, R (i.e., $R_1 \div R_2$) contains set of all tuples t such that for any tuple t_1 and t_2 of R_1 and R_2 respectively, following conditions are hold.

- $t[P - Q] = t_1[P - Q]$
- $t_1[P - Q] = t_2[Q]$

2.2 Reduct Generation

The measurement of similarity/dissimilarity among the genes based on the distance metric may not be effective for gene data analysis in a high dimensional space. And at the same time, elegant sample selection decreases the workload and simplifies the subsequent design process to a great extent. So, the method proposed a design approach to compute a minimum subset of samples called reduct which can, by itself, fully characterize the knowledge in the gene database as the whole set of attributes (C) and preserves partition of data. It generates initial reduct RED by selecting the samples with score less than the average score. But in most of the cases, this initial reduct could not fully characterized the knowledge and so other samples are added to RED one at a time based on their score only if resultant samples more accurately characterized the knowledge than that obtained previously. To obtain the final reduct, initially, gene dataset is partitioned into k clusters, say, $CLASS = \{CLAS_1, CLAS_2, \dots, CLAS_k\}$ and the dataset reduced by RED is clustered and validated by [7] and obtain say, l clusters of genes namely, $CLUS = \{CLUS_1, CLUS_2, \dots, CLUS_l\}$. Now, a matching factor of CLUS to CLASS is calculated for RED, using equation (3) which implies how much maximum number of genes in clusters is correctly positioned based on their class labels.

$$mf_{RED}(CLUS) = \sum_{j=1}^k \frac{1}{|CLUS_j|} \max_{1 \leq i \leq l} \{ \{CLUS_j \cap CLASS_i\} \} \quad (3)$$

Finally, samples not in RED are arranged in increasing order of their score in NRED and examined separately to determine if they are to be included in the final reduct set. So, each sample C_i in NRED is added to RED and apply same clustering algorithm and then compute matching factor using (3). If the matching factor is larger than the previously obtained matching factor then the sample is kept in RED which implies that resultant samples more accurately characterized the knowledge than that obtained previously; otherwise the sample C_i is discarded from RED. Repeating the process for all samples in NRED, final reduct is obtained.

Algorithm: Reduct_Generation(DS, RED)

/*DS = (U, C, D), where C is the number of samples and D is the decision attribute with k distinct values and RED is the final reduct of samples*/

Begin

RED = \emptyset ;

For each sample $C_i \in C$

```

    Compute Score  $S(C_i)$  using equation (2);
    Compute average score avg_sc;
    For each sample  $C_i \in C$  /*initial reduct formation*/
        If  $(S(C_i) < \text{avg\_sc})$  RED = RED  $\cup$   $\{C_i\}$ ;
    Genes in set U are partitioned into k classes based
    on their D-values;
    Let CLASS =  $\{CLAS_1, CLAS_2, \dots, CLAS_k\}$ ;
    NRED = C - RED;
    Let matching factor  $mf = -\infty$ ;
    Repeat { /* Final reduct formation*/
        Genes in U are clustered into l groups based on
        RED by clustering algorithm[7];
        Let CLUS =  $\{CLUS_1, CLUS_2, \dots, CLUS_l\}$ ;
        Compute  $mf_{RED}(CLUS)$  of CLUS for RED by equation(3);
        If  $(mf_{RED}(CLUS) > mf)$   $mf = mf_{RED}(CLUS)$ 
        Else RED = RED -  $\{C_t\}$ 
        Remove lowest score sample  $C_t$  of NRED,
        /*i.e., NRED = NRED -  $\{C_t\}$ */
        RED = RED  $\cup$   $\{C_t\}$ ;
    } Until (NRED =  $\phi$ );
End.

```

3 Experimental Results and Performance Evaluation

Experimental studies presented here provide an evidence of effectiveness of proposed dimension reduction technique. The six microarray gene dataset used in the experiment are described below where each row in the data sets represents the expression pattern of one gene and each column represents an experimental sample. Experiments were carried out on large number of different kinds of microarray data (cancerous data) few of them described below are summarized. Each dataset contains two types of samples, one group is normal and other is cancerous.

(i) Colon Cancer dataset: The colon cancer data contains 2000 out of around 6500 genes and 62 samples collected from colon-cancer patients. The raw data are available at <http://microarray.princeton.edu/oncology/affydata/index.html>.

(ii) Diffuse Large B-cell Lymphoma (DLBCL) dataset: There are two kinds of classifications about DLBCL versus Follicular Lymphoma (FL) morphology. This set of data contains 58 DLBCL samples and 19 FL samples. The expression profile contains 7129 genes. Raw data are available at <http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi>.

(iii) Leukemia (ALL V.S. AML) dataset: Training dataset consists of 38 bone marrow samples (27 ALL and 11 AML), over 7129 human genes. The raw data are available at <http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi>.

(iv) Prostate Cancer dataset: Test set contains 21 samples and 13 patients having remained relapse free prostate samples with around 12600 genes. The raw data are available at <http://www-genome.wi.mit.edu/mpr/prostate>.

(v) **Central Nervous (C. N.) System dataset:** The data set contains 60 patient samples, 21 are survivors and 39 are failures. There are 7129 genes in the dataset. The raw data are available at <http://www-genome.wi.mit.edu/mpr/CNS>.

(vi) **GDS2771 series data:** There exist 2000 genes and 72 samples. Among these, 36 samples are normal and others are cancerous. The original data are available at <https://www-r-forge.r-project.org/R>.

The proposed technique generates final reducts for above six datasets. The method is compared with well known dimension reduction algorithms such as Principal Component Analysis (PCA), Singular Value Decomposition (SVD) and Consistency Subset Evaluation (CSE) by evaluating DB-index (Davies-Bouldin index) [10] and RMS (Root Mean Square) error [11] in cluster sets. The datasets reduced by the proposed and above dimension reduction method are partitioned by clustering algorithm [7]. Then DB index and RMS error are computed for the clusters and listed in Table 1. Smaller the DB-index [10] and RMS error [11] better is the clustering. The Proposed, PCA and SVD and CSE algorithms are implemented using Mat lab 7.8.1 version. The results show that DB index and RMS errors produced by the proposed method are less than that produced by other methods for all cancerous microarray data sets, which confirms the potentiality and superiority of the proposed method.

Table 1. DB index and RMS error for proposed and other methods

Data Name	Methods							
	PCA		SVD		CSE		Proposed	
	DB index	RMS error	DB index	RMS error	DB index	RMS error	DB index	RMS error
Colon Cancer	0.0665	124.89	0.0814	290.98	0.0611	158.95	0.04048	88.071
DLBCL	0.0406	243.63	0.0259	258.98	0.0235	120.59	0.02145	229.60
ALL VS. AML	0.0336	167.03	0.0394	202.07	0.0315	157.64	0.02906	152.87
Prostate Cancer	0.0429	110.85	0.0515	133.91	0.0614	174.41	0.03913	94.828
C.N. System	0.1102	1236.9	0.1926	5134.1	0.1471	1266.5	0.09970	1128.1
GDS2771 series	0.0722	370.96	0.1416	530.47	0.1103	951.98	0.04350	272.23

The original dataset and the datasets reduced by proposed and other dimension reduction methods are classified by the classification methods [12-13] such as Bayes classifier (Naïve Bayes), Trees based classifier (J48-C 0.25), Rules based classifier (PART), Functions based classifier (MultiLayerPerceptron), Trees based classifier (RandomForest), Meta classifier (Bagging), and Lazy classifier (Kstar) and accuracies are plotted with various colours, as shown in Fig. 1 to Fig. 6. It is observed that for almost all datasets proposed method shows better accuracy for most of the classifiers. All classification performances are measured by Weaka-3-6-5 Data Mining tool and comparison figures are drawn in Matlab 7.8.1 version.

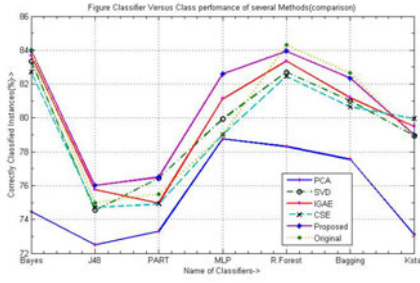


Fig. 1. Colon Cancer Dataset

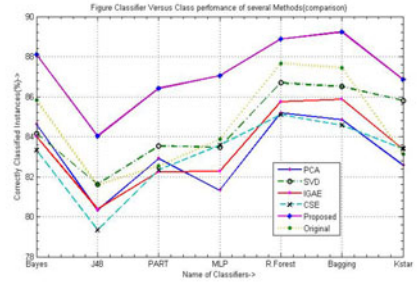


Fig. 2. DLBCL Dataset

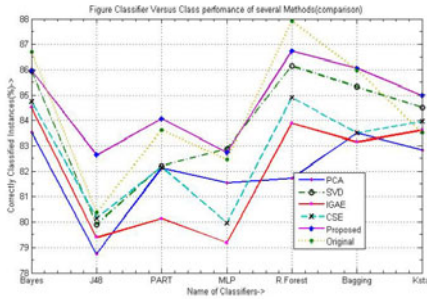


Fig. 3. Leukemia Dataset

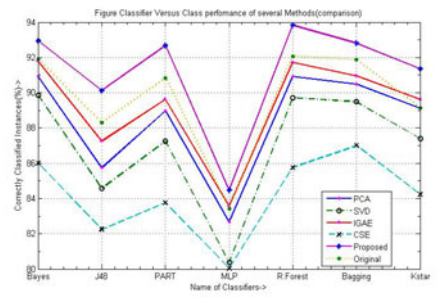


Fig. 4. Prostate Cancer Dataset

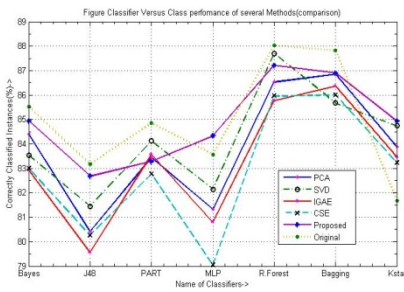


Fig. 5. Central Nervous System Dataset

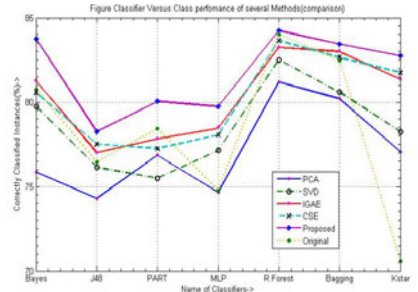


Fig. 6. GDS2771 Series Dataset

4 Discussions and Conclusions

In the paper a novel dimension reduction technique has been proposed for obtaining suitable number of samples based on the concept of division operation of relational algebra in database management system. Initially gene expression microarray dataset is discretized and labeled by some meaningful linguistic variables and finally reduct is generated. Experimental results shown for seven different kinds of microarray

cancerous data evaluates the performance of the proposed algorithm both qualitatively as well as quantitatively. Comparative study is made with traditional dimension reduction algorithms namely PCA, SVD, Info Gain Attribute Evaluation (IGAE) and Consistency Subset Evaluation (CSE) with respect to DB index and Root Mean Square error (RMS) which shows that the proposed method selects fairly well dimensions in terms of the clustering quality. Comparative study is also made with same reduction algorithms with respect to correctly classified instances (%) by some traditional classifiers namely Bayes, J48, PART, MLP, Random Forest, Bagging and Kstar which shows the goodness of proposed method. Gene identification for cancer detection is the future work based on the reduced samples obtained by the proposed work.

References

1. Aerman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J.: Expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci.* 1, 6745–6750 (1999)
2. Hand, D.J., Heard, N.A.: Finding groups in gene expression data. *Journal of Biomedicine and Biotechnology* 2, 215–225 (2005)
3. Muralidhar, K., Sarathy, R.: Security of random data perturbation methods. *ACM Trans. Database Syst.* 24(4), 487–493 (1999)
4. Petrov, A., Shams, S.: Microarray image processing and quality control. *VLSI Signal Processing* 38(3), 211–226 (2004)
5. Siedlecki, W., Sklansky, J.: On automatic feature selection. *International Journal of Pattern Recognition and Artificial Intelligence* 2(2), 197–220 (1988)
6. Ding, C., Peng, H.C.: Minimum Redundancy Feature Selection from Microarray Gene Expression Data. In: *Proc. Second IEEE Computational Systems Bioinformatics Conf.*, pp. 523–528 (2004)
7. Pati Kr., S., Das Kr., A.: Cluster Analysis of Microarray Data Based on Singularity Measurement. *International Journal of Bioinformatics Research* 3(2), 207–213 (2011) ISSN: 0975-3087
8. Silberschatz, A.: *Introduction to Data base Management System*. Tata McGraw Hill, New Delhi
9. Garey, M., Johnson, D.: *Computers and intractability: A guide to the theory of NP-completeness*. Freeman, New York (1979)
10. Davies, L., Bouldin: Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1(2), 95–104 (1979)
11. Huffman George, J.: Estimates of Root-Mean-Square Random Error for Finite Samples of Estimated Precipitation, pp. 1191–1201. *American Meteorological Society* (1997)
12. Jirapech-Umpai, T., Aitken, S.: Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. *BMC Bioinformatics* 6(148) (2005)
13. Nguyen, D.V., Rocke, D.M.: Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* 18(1), 39–50 (2002)
14. Huynen, M., Snel, B., Lathe III, W., Bork, P.: *Genome Res.* 10, 1204–1210 (2000)
15. Mollr-Levet, C., Cho, S., Wolkenhauer, O.: Microarray data clustering based on temporal variation: Fcv and tsd preclustering. *Applied Bioinformatics* 2(1), 35–45 (2003)

Hybrid PSO - Bacterial Foraging Based Intelligent PI Controller Tuning for pH Process

G. Petchinathan, G. Saravanakumar, K. Valarmathi, and D. Devaraj

Kalasalingam Academy of Research and Education,
Krishnankoil, Tamilnadu, India
gpetchi@gmail.com

Abstract. The control of pH process is a difficult problem due to its inherent nonlinearity and time-varying characteristics. For the pH process, Proportional Integral (PI) control has been successfully used for many years. Tuning of the PI controller is necessary for the satisfactory operation of the system. This paper proposes a hybrid approach involving Bacterial Foraging Optimization (BFO) Algorithm and Particle Swarm Optimization (PSO) algorithm for determining the optimal proportional-Integral (PI) controller parameters for control of a pH Process. The BFO algorithm depends on random search directions which may lead to delay in reaching the global solution. The PSO may lead to possible entrapment in local minimum solutions. The proposed hybrid approach has stable convergence characteristic and good computational efficiency. Simulation results clearly illustrate that the proposed approach is very efficient in improving the step response characteristics such as, reducing the Mean Square Error (MSE), rise time and settling in control of a pH process.

Keywords: pH Process, PI controller, PSO-BFO algorithm, Mean Square Error, Settling Time.

1 Introduction

Over the last 50 years, many ways have been developed to determine PI controller parameters for stable processes suitable for auto tuning and adaptive control [1–4]. Such tuning uses only a small amount of information about the system's dynamic behavior and often does not provide good tuning. Ant Colony Optimization (ACO) was introduced around 1991-1992 by M. Dorigo and colleagues as a novel nature-inspired metaheuristic for the solution of hard combinatorial optimization problems [5]. Farooq et al [6] developed a bee inspired algorithm for routing in telecommunication network. Swarming strategies in bird flocking and fish schooling are used in the Particle Swarm Optimization (PSO) introduced by Eberhart and Kennedy and it is easy to implement and there are few parameters to adjust and this algorithm has been successfully applied in many areas [7]. A relatively newer evolutionary computation algorithm called Bacterial Foraging scheme has been proposed and introduced recently by K. M. Passino [8]. This algorithm inspired by the behavior of *Escherichia Coli* (*E. Coli*) bacteria normally lives inside the intestines where it helps the body to break down and digest the food. In this paper, the use of

both PSO and E. coli based BFO algorithms for tuning of PI controller is investigated in control of pH process. A proposed approach that combines the above mentioned optimization algorithms.

2 pH Process Modeling

The pH is the measurement of the acidity or alkalinity of a solution. The pH process consists of neutralization of two monoprotic reagents of a weak acid and a strong base. The model of the pH neutralization process used in this work follows that proposed by McAvoy et al. [9], [10] and is shown in Fig. 1. Assumption of perfect mixing is general in the modeling of pH processes. Material balances in the reactor can be given by

$$V \frac{dx_a}{dt} = F_a C_a - (F_a + F_b) X_a \tag{1}$$

$$V \frac{dx_b}{dt} = F_b C_b - (F_a + F_b) X_b \tag{2}$$

Where F_a is the flow rate of the influent stream, F_b is the flow rate of the titrating stream, C_a is the concentration of the influent stream, C_b is the concentration of the titrating stream, x_a is the concentration of the acid solution, x_b is the concentration of the basic solution and V is the volume of the mixture in the CSTR.

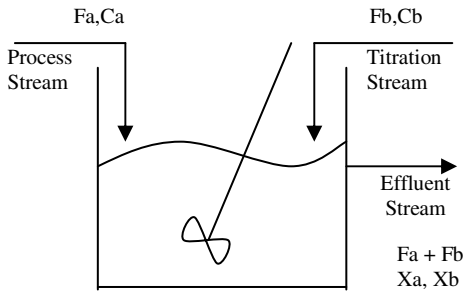


Fig. 1. pH neutralization Process

3 Particle Swarm Optimization (PSO)

The PSO method is a member of wide category of Swarm Intelligence methods for solving the optimization problems. It is a population based search algorithm where each individual is referred to as particle and represents a candidate solution [7], [11]. In PSO each particles strive to improve themselves by imitating traits from their successful peers. [12]. Each particle has a position represented by a position-vector X_k^i where (i is the index of the particle), and a velocity represented by a velocity

vector V_k^i . Each particle remembers its own best position P_{Lbest}^i . The best position vector among the swarm then stored in a vector P_{Global}^i . During the iteration time k , the update of the velocity from the previous velocity to the new velocity is determined by.

$$V_{k+1}^i = V_k^i + C_1 R_1 (P_{Lbest}^i - X_k^i) + C_2 R_2 (P_{Global}^i - X_k^i) \tag{3}$$

The new position is then determined by the sum of the previous position and the new velocity.

$$X_{k+1}^i = X_k^i + V_{k+1}^i \tag{4}$$

A particle decides where to move next, considering its own experience, which is the memory of its best past position, and the experience of the most successful particle in the swarm.

4 Bacterial Foraging Optimization (BFO)

In foraging theory, it is assumed that the objective of the animals is to search for and obtain nutrients in such a fashion that the energy intake per unit time is maximized [12]. The foraging behavior of E. coli bacteria present in our intestines, which includes the methods of locating, handling and ingesting food, has been successfully mimicked to propose a new evolutionary optimization algorithm [8] ,[13]. This optimization procedure comprises of four basic steps: a) chemotaxis, b) swarming, c) reproduction and d) elimination and dispersal. The objective will be to try and implement a biased random walk for each bacterium where it will try to climb up the nutrient concentration and try and avoid noxious substances and will attempt to leave a neutral environment as soon as possible.

5 PSO Based Bacterial Foraging Optimization (PSO-BFO)

PSO based BFO combines both PSO and BFO algorithms [14]. This combination aims to make use of PSO ability to exchange social information and BFO ability in finding a new solution by elimination and dispersal.

For initialization, the user selects $S, N_s, N_c, N_{re}, N_{ed}, P_{ed}, C_1, C_2, R_1, R_2$ and $c(i), i=1, 2, \dots, S$. Also initialize the position $P_n^i, i=1, 2, \dots, S$. and velocity randomly initialized. Fig.2 shows the flow chart for PSO based BFO algorithm. Initially, $j=k=ell=0$ and Initialize parameters $n, S, N_s, N_c, N_{re}, N_{ed}, P_{ed}, C_1, C_2, R_1, R_2$ and $c(i), i=1, 2, \dots, S$. and Delta as shown in table 2.

6 PI Controller Tuning by BFO, PSO and PSO-BFO

The controller tuning is actually required to reduce the Mean Square Error, overshoot and settling time in step response of the pH process. Hence by using the proportional and integral (PI) controllers above stated can be easily achieved.

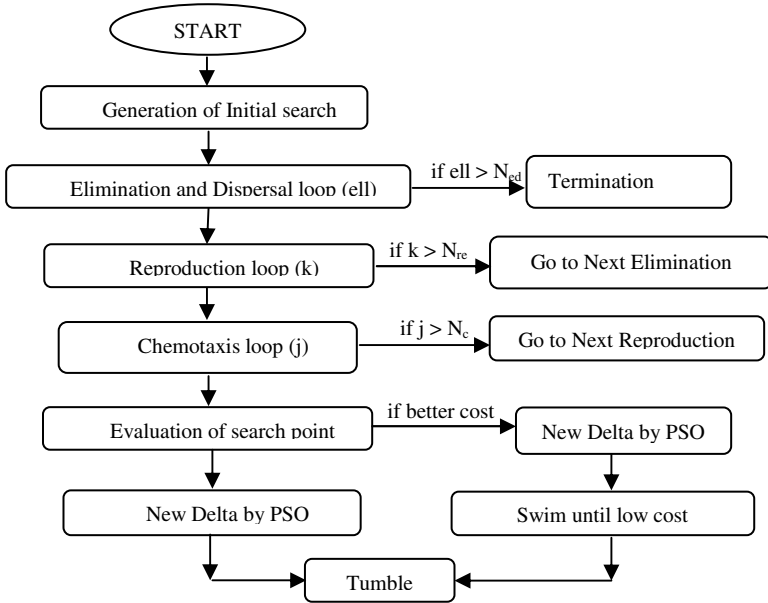


Fig. 2. Flow chart for PSO based BFO Algorithm

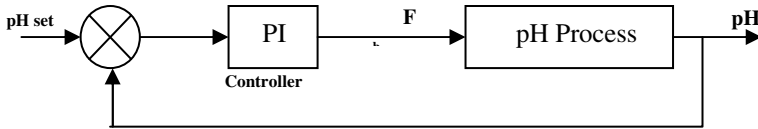


Fig. 3. Block Diagram of a pH process with PI controllers

Attempt has been made to achieve globally minimal Mean square error criteria in the step response of a pH process which is cascaded with PID controller by tuning the K_p proportional gain and K_i integral gain values. In the transfer function of the controller stated as

$$G_c(s) = K_p + (K_i/s) \tag{5}$$

Fig. 3 shows closed loop of pH process with PI controller. This section presents the details of the design of PI controller for the pH process. First the GA, BFO and PSO algorithms were applied to design the PI controller for a simulated pH process. Then proposed hybrid PSO based BFO algorithm was applied to design the PI controller for a simulated pH process. The implementation of this hybrid algorithm was written in MATLAB and executed on a PC with Pentium duo core processor. The pH process was simulated based on the Equations (1) and (2) using MATLAB simulink with the model parameters for the experimental system which is tabulated in Table.3. Fig.4 shows simulink diagram for step response of pH model with PI controller. Step signal is used as the input to the system.

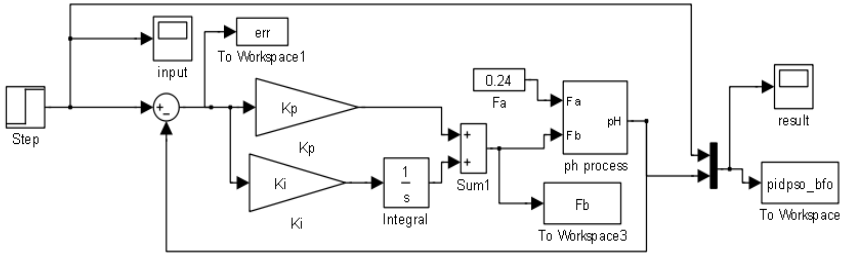


Fig. 4. Simulink diagram for step response of pH process with PI controller

The GA, BFO, PSO and PSO-BFO algorithms are implemented to find the optimal parameters of the controller. The output of PI controller was used to control the pH process by manipulating the base flow (F_b) and acid flow (F_a) is kept constant. The objective function in control of pH process is to minimization of Mean square error (MSE). The performance of the GA, BFO, PSO and PSO-BFO algorithms are evaluated with constant values of initial parameters mentioned in the table 2.

Table 1. Comparison of different tuning methods

pH	Fa	Type	Kp	Ki	MSE
5	0.192	GA	9	8	$3.75e^{-2}$
		BFO	8.8289	3.7389	$4.30e^{-3}$
		PSO	72.972	11.525	$2.36e^{-2}$
		PSO - BFO	103.51	4.3212	$2.87e^{-3}$
	0.288	GA	9.677	9.01	$4.51e^{-2}$
		BFO	8.791	9.653	$4.87e^{-3}$
		PSO	134.2	24.478	$3.574e^{-2}$
7	0.192	GA	1.1624	11.764	$4.65e^{-2}$
		BFO	5.411	1.38	$1.714e^{-2}$
		PSO	5.6113	53.329	$2.97e^{-2}$
		PSO - BFO	0.2996	1.9352	$2.536e^{-3}$
	0.288	GA	1.129	11.764	$1.038e^{-2}$
		BFO	2.97	7.45	$2.26e^{-3}$
		PSO	1.7490	0.7217	$2.13e^{-3}$
11	0.192	PSO - BFO	0.2743	2.0024	$1.95e^{-3}$
		GA	12	9	$1.05e^{-1}$
		BFO	4.225	2.202	$2.53e^{-1}$
		PSO	267.93	68.544	$8.09e^{-2}$
	0.288	PSO - BFO	38.37	0.0573	$7.316e^{-2}$
		GA	12	8.5405	$1.06e^{-1}$
		BFO	6.002	7.018	$2.615e^{-1}$
0.288	PSO	50.5877	0.56357	$9.354e^{-2}$	
	PSO - BFO	44.542	0.1826	$4.507e^{-2}$	

Table 2. Parameters value of each algorithms

Parameter	Symbol	PSO – BFO	BFO	PSO	GA
No. of bacteria in the population	S	10	50	50	10
Dimension of search space	n	2	2	2	--
Maximum no. of swim length	N _s	4	4	--	--
No. of Chemotatic steps	N _c	20	50	50	30
No. of reproductive steps	N _{re}	2	2	--	--
No. of elimination dispersal events	N _{ed}	2	2	--	--
Elimination dispersal probability	P _{ed}	0.25	0.25	--	--
Step size	C(i)	0.5	0.05	--	--
Cognitive factor	C ₁	1.2	--	1.2	--
Social acceleration factors	C ₂	0.5	--	0.12	--
Momentum/ Inertia	w	0.9	--	0.9	--
Cross over probability	P _c	--	--	--	0.8
Mutation probability	P _m	--	--	--	0.08

Table 3. Model parameters for the pH process

Parameter	Description	Value
<i>V</i>	Volume of the Continuous Stirred Tank Reactor	7.4l lit
<i>F_a</i>	Flow rate of the influent stream	0.24 l min-1
<i>F_b</i>	Flow rate of the titrating stream	0-0.8 l min-1
<i>C_a</i>	Concentration of the influent stream	0.2 g mol l-1
<i>C_b</i>	Concentration of the titrating stream	0.1 g mol l-1

In order to examine the effectiveness of the proposed control scheme, simulations are carried out in different cases, e.g. in different set points of pH value (5, 7 and 11) and for different influent stream flow rate (*F_a*). For simulation *F_a* value is taken as 20 percent deviation from the value mention in the table 3 (*F_a*=0.192 and 0.288). In this paper, the results of the proposed hybrid PSO – BFO algorithm is compared with the results of a GA, BFO and PSO algorithms as shown in table1. Fig. 5, Fig.7 and Fig. 9 shows the step response of pH process with acid flow rate *F_a*=0.192 for set point of pH=5,7and 11 respectively. Fig. 6, Fig.8 and Fig. 10 shows the step response of pH process with acid flow rate *F_a*=0.288 for set point of pH=5,7and11 respectively. In all the cases of step response PSO-BFO results s lower value of MSE and settling time compared with other optimization methods. Overshoot in step response for all the cases are almost zero.

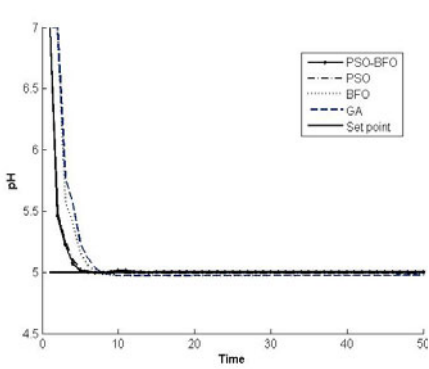


Fig. 5. Step response of pH process for a step input pH = 5 and $F_a=0.192$

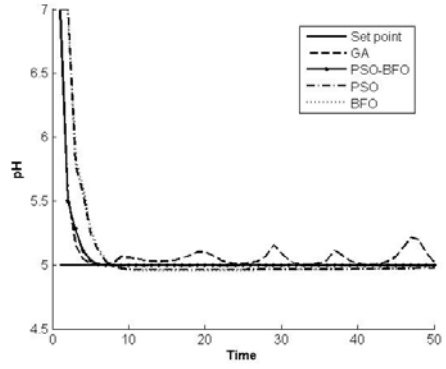


Fig. 6. Step response of pH process for a step input pH = 5 and $F_a=0.288$

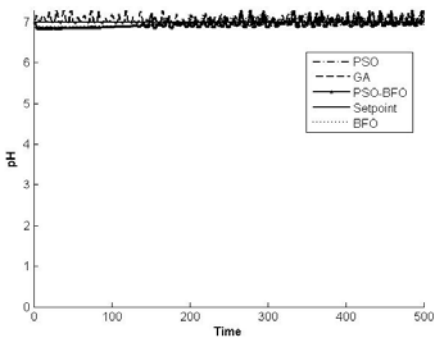


Fig. 7. Step response of pH process for a step input pH = 7 and $F_a=0.192$

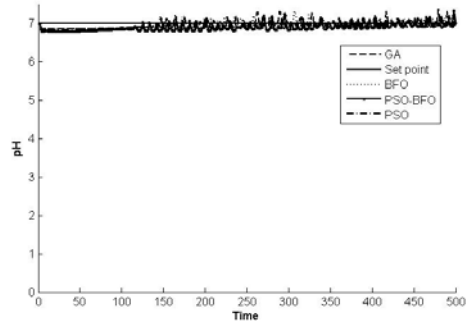


Fig. 8. Step response of pH process for a step input pH = 7 and $F_a=0.288$

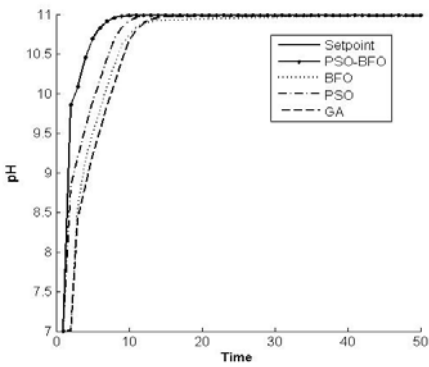


Fig. 9. Step response of pH process for a step input pH = 11 and $F_a=0.192$

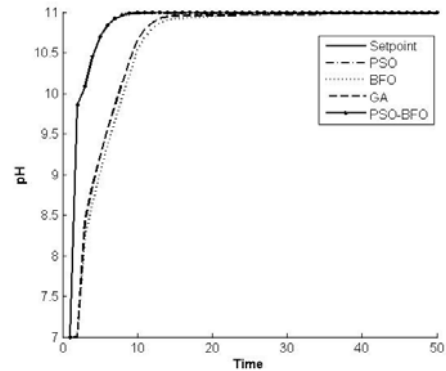


Fig. 10. Step response of pH process for a step input pH = 11 and $F_a=0.288$

7 Conclusion

In this proposed work, the optimal parameters of the PI controller at each pH region are computed by using hybrid PSO based BFO algorithm. This algorithm combines PSO and BFO techniques to make use of exchange social information ability of PSO and elimination and dispersal ability of BFO in finding a new solution. From the simulation results (Fig. 5-10 and Table. 1) the PSO based BFO tuned PI controller has minimum MSE and settling time compared with BFO, PSO and GA.

References

1. ACPA, Special Issue on Advance in PID control. Asian Journal of Control 4(4) (2002)
2. Astrom, K.J., Hagglund, T.: The Future of PID Control. IFAC J. Control Engineering Practice 9, 1163–1175 (2001)
3. Xu, J.X., Liu, C., Hang, C.C.: Tuning of Fuzzy PI Controllers Based on Gain/Phase Margin Specifications and ITAE Index. ISA Transactions 35, 79–91 (1996)
4. Zhung, M., Atherton, D.P.: Automatic Tuning of Optimum PID Controllers. IEE Proc. D, Control Theory and Applications 140(3), 216–224 (1993)
5. Dorigo, Blum: Ant Colony Optimization Theory: A survey. TCS: Theoretical Computer Science 345 (2005)
6. Farooq, M.: From the Wisdom of the Hive to Intelligent Routing in Telecommunication Networks, February 01 (2006)
7. Kennedy, J., Eberhart, R.C.: Particle Swarm Optimization. In: Proc. of the IEEE Int. Conf. on Neural Networks, pp. 1942–1948. IEEE Service Center, Piscataway (1995)
8. Passino, K.M.: Biomimicry of Bacterial Foraging for Distributed Optimization and Control. IEEE Control Systems Magazine 22(3), 52–67 (2002)
9. McAvoy, T.J., Hsu, E.: Dynamics of pH in controlled stirred tank reactor. Ind. Eng. Chem. Process Des. Dev. 68, 114–120 (1972)
10. Valarmathi, K., Devaraj, D., Radhakrishnan, T.K.: Adaptive Enhanced Genetic Algorithm-Based Proportional Integral Controller Tuning for pH Process. Taylor and Francis-Instrumentation Science & Technology 35(6), 619–635 (2007)
11. Eberhart, R.C.: Computational Intelligence: A Perspective. In: Evolutionary Programming, pp. 239–245 (1996)
12. Panda, S., Padhy, N.P.: Comparison of Particle Swarm Optimization and Genetic Algorithm for TCSC-based Controller Design. Int. Journal of Computer Science and Engineering 1(1), 41–49 (2007)
13. Kim, D.H., Cho, J.H.: Robust Tuning of PID Controller Using Bacterial-Foraging-Based Optimization. Journal of Advanced Computational Intelligence and Intelligent Informatics 9(6), 669–676 (2005)
14. Korani, W.M.: Bacterial Foraging Oriented by Particle Swarm Optimization Strategy for PID Tuning. In: Proceedings of the GECCO Conference Companion on Genetic and Evolutionary Computation (2008) ISBN: 978-1-60558-131-6

Grid Computing Based NIC Infrastructure: A Step towards IT Enabled India

Buddhadeb Pradhan, Rabindra Kumar Shial, and Diptendu Sinha Roy

National Institute of Science and Technology, Berhampur, India
{Buddhadebpradhan, diptendu.sr}@gmail.com,
rkshial@yahoo.com

Abstract. Public administration is administered by bureaucratic constitution and is built on rationale principles. Though dominant during the twentieth century, such system has failed to react to the shifting requirements of the current times. E-governance, which is a paradigm shift over the traditional approaches in public administration, means rendering of government services and information to the public using electronic means. This new paradigm has revolutionized the quality of service delivered to the citizens. It has ushered in transparency and simplicity in the governing process; enormous time saving due to stipulation of services through single window; simplification of procedures; better office and record management; reduction in corruption; and improved attitude, behavior and job handling capacity of the dealing personnel. With the proliferation of Information and Communication Technologies (ICTs), the existent information services catering e-governance facilities to citizens needs a technology change in order to provide the required seamless services. This paper presents the case of suitability of Grid Computing to the already existing National Informatics Center (NIC) infrastructure of India.

1 Introduction

Previously, service delivery mechanisms of the government departments left much to be desired in India. restricted spaces; untidy ambience; ill-mannered dealing personnel and their constant absenteeism; demands of indulgence; inefficiency in work; long queues; procrastinating officials; procedural complexities; etc., were some of the undesirable features of the working of the government departments. As a result, a visit to government department by a citizen to make use of any service used to be a traumatic experience. With the increasing consciousness amongst the citizens and their better experiences with the private sector – the demand for better services on the part of government departments became more pronounced. The combination of Information and Communication Technology (ICT) has played a prominent role in strengthening such a demand. The metamorphosis in the quality of delivery of services to the citizens by the government has been more pronounced in recent years with the advent of e-governance [1]. E-governance, which is a archetype shift over the traditional approaches in Public Administration, means representation of government services and information to the public using electronic means. Basic architecture of E-governance has been given in figure 1.

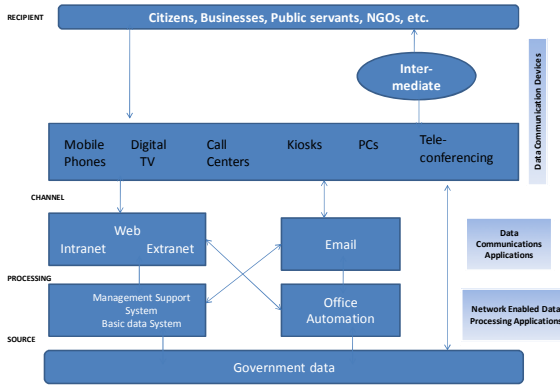


Fig. 1. Architecture of E-governance

This new prototype has brought about a revolution in the quality of service delivered to the citizens. It has ushered in transparency in the governing process; saving of time due to provision of services through single window; simplification of procedures; better office and record management; reduction in corruption; and improved attitude, behavior and job handling capacity of the dealing personnel. The present study substantiates these theoretical assumptions about e-governance by analyzing some experiences at the Union as well as State Government Level in India [2]. E-governance state wise scheme is given in Table 1 (Courtesy [6]).

Table 1. State wise list of e-government schemes in India

State/Union Territory	Initiatives covering departmental automation, user charge collection, delivery of policy/programme information and delivery of entitlements
Andhra Pradesh	e-Seva, CARD, VOICE, MPHS, FAST, e- Cops, AP online – one –stop-shop on the internet, Saukaryam, Online transaction processing, e-immunization Rural Health Call Center and Site Suitability for Water Harvesting, Professional e-Pension
Bihar	Sales Tax Administration Management Information, E-Khajana
Chhattisgarh	Chhattisgarh InfoTech Promotion Society, Treasury Office, e-linking project
Delhi	Automatic Vehicle Tracking System, Computerization of website of RCS office, Electronic clearance system, Management Information System of Education, Delhi Slum Computer Kiosks etc.
Goa	Dharani Project
Gujarat	Mahiti Shakti, Dairy Information System Kiosk (DISK), Request for government documents online, Form Book Online, G R book Online, Census Online, Tender Notice.
Haryana	Nai Disha, Result through Binocular
Himachal Pradesh	Lok Mitra, HIMRIS ,e-pension, Unreserved Ticketing System by Indian Railways
Jharkhand	Vahan, Tender Notice
Karnataka	Bhoomi, Kaveri, Khazane
Kerala	e-Srinkhla, RDNet, Fast, Reliable, Instant, Efficient Network for the Disbursement of Services (FRIENDS)

Table 1. (continued)

Madhya Pradesh	Gyandoot, Gram Sampark, Smart Card in Transportation Department, Computerization MP State Agricultural Marketing Board (Mandi Board), Headstart etc.
Maharashtra	SETU, Koshvani, Warana Wired Villages, Telemedicine Project (Pune), Online Complaint Management System Mumbai
Odisha	E-Shishu, Common service centres (CSCs) in panchayats
Punjab	SUWIDHA(Single User Window Disposal Help Line for Applicants), SUBS(SUwidha Backend Services), AGMARKNET(Agriculture Marketing Network), ALIS(Arms License Information System), TISP(Treasuries Information System of Punjab), SSIS(Social Security Information System), WEBPASS(District Passport Application Collection Centre)
Rajasthan	Jan Mitra, RajSWIFT, Lokmitra, RajNIDHI, Aarakshi - Online FIR, Professional EDelivery of Tax Payers by Income Tax
TamilNadu	Rasi Maiyama-Kanchipuram, Application Forms Related to Public Utility, Tender Notice & Display
Uttar Pradesh	Lokvani, e Suvidha, Bhulekh, (Land Records), Koshvaani, Treasury Computerization, PRERNA: PROperty Evaluation and Registration Application, Bouquets of services offered by Transport Department
Uttarakhand	Kisan Soochna Kutirs (KSKs) , Village Information Centres (VICs), Computerization of Land Record Department, Automation of Transport Department:
West Bengal	Vehicle registration, land records, birth and death registrations, employment exchanges, payment of excise duty, sales tax and local tax, electronic bill payment of water and electricity, computerization of health records.
North Eastern State	
Assam	ASHA
Arunachal Pradesh, Manipur, Meghalaya, Mizoram & Nagaland	Community Information Centre. Forms available on the Meghalaya website under schemes related to social welfare, food civil supplies and consumer affairs, housing transport etc.

No doubt, India has introduced these global trends/ measures in 1990, but no sincere exercise has been undertaken in the corresponding 15 years to examine the effects of these the role of the information technology, in the governance process. The present paper is an attempt to fill this gap in the existing literature. The term governance needs to be understood before we move on to e-government and e-governance [3].

The two terms- e-government and e-governance are independent of each other, but are at times used alternatively, there by the major distinction between e-government and e-governance is missed out. E-government is understood as the use of Information and Communication Technology (ICT) to promote more efficient and cost effective government, facilitate more convenient government services and allow greater public access to information, and make government more accountable to citizens, where as governance is a wider term which covers the state's institutional arrangements, decision making processes, implementation capacity and the relationship between government officials and the public. E-governance is the use of ICT by the government, civil society and political institutions to engage citizens through dialogue and feedback to promote their greater participation in the process of governance of these institutions [4]. Thus, e-government can be viewed as a subset of

e-governance, and its focus is largely on improving administrative efficiency and reducing administrative corruption (Bhatnagar Subhash, 2004). Overall organization structure of NIC has been depicted on figure 2(Courtesy [4]).

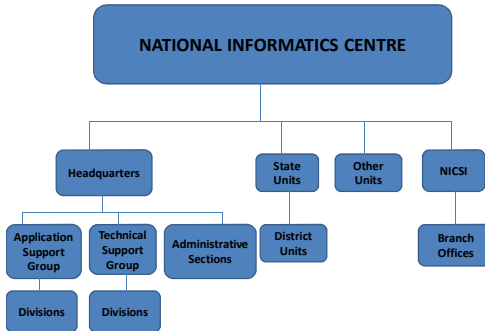


Fig. 2. Organization Structure of NIC

1.1 Scope of E-Government

While e-government encompasses a wide range of activities, we can identify three distinct areas. These include government-to-government (G to G), government-to-citizens (G to C), and government to business (G to B). Each of these represents a different combination of motivating forces. However, some common goals include improving the efficiency, reliability, and quality of services for the respective groups. In many respects, the government to government (G to G) sector represents the backbone of e-government and it involves sharing data and conducting electronic exchanges between various governmental agencies. One benefit with this is cost savings, which is achieved by increasing the speed of the transactions, reduction in the number of personnel necessary to complete a task, and improving the consistency of outcomes [5]. Government to citizen (G to C) facilitates citizen interaction with government, which is primary goal of e-government. This attempts to make transactions, such as payment of taxes, renewing licenses and applying for certain benefits, less time consuming and easy to carry out [5].

1.2 E-Government Initiatives in India: An Overview

The Government of India kick started the use of IT in the government in the right earnest by launching number of initiatives. First the Government approved the National E-Governance Action plan for implementation during the year 2003-2007. The plan is an attempt to lay the foundation and provide impetus for long-term growth of e-governance within the country. It proposed to create the right governance and institutional mechanisms at the center, state and local levels to provide a citizen centric and business centric environment for governance. Apart from the action plan, the following measures have also been introduced [6]:

Adoption of “Information Technology (IT) Act, 2000 has been introduced by the Government of India to provide legal framework to facilitate electronic transactions. The major aims of this act are to: recognize electronic contracts, prevents computer crimes, and make electronic filing possible. The Act came into force on 17 October, 2000; establishment of the National Taskforce of Information Technology and Software Development in May 1998, creation of Centre for e-governance to disseminate the best practices in the area of e- governance for the use by the Central and State Governments and act as a nodal center to provide general information on e-governance, national and international initiatives, and IT policies of the government(s) and developing e-office solutions to enable various ministries and departments to do their work electronically. Modules such as Workflow for Drafts for Approvals, e-file, e-noting, and submission of reports, integrated personal information and financial accounting systems have been developed [7].

This remainder of this paper has been organized as follows: Section 2 presents the Role of NIC’s in India and discusses about its present and future scope. Section 3 describes the distributed nature of grid computing and its suitability for distributed management system. Section 4 gives the details about the proposed grid system for NIC and paper concludes at Section 5.

2 Role of NIC’s in India: Present and Future Scope

NIC, under the Department of Information Technology of the Government of India, is a premier Science and Technology organization, at the forefront of the active promotion and implementation of Information and Communication Technology (ICT) solutions in the government. Social, geographical and economical disparity issues have to be removed and proper infrastructure is required to establish e-governance. The ICT facilities need to be developed and should be available to one and all citizenry. Internet connection through satellite, phone lines or through cable or Television should be accessible for all specially to the people in rural areas [8].

As a major step in ushering in e-Governance, NIC implements the following minimum agenda as announced by the Central Government:

Internet/Intranet Infrastructure (PCs, Office Productivity Tools, Portals on Business of Allocation and Office Procedures), IT empowerment of officers/officials through Training, IT enabled Services including G2G, G2B, G2C, G2E portals , IT Plans for Sector Development, Business Process Re-engineering, NIC provides a rich and varied range of ICT services delineated below.

Digital Archiving and Management, Digital Library, E-Commerce, E-Governance, Geographical Information System, IT Training for Government Employees and so forth [9].

NIC endeavours to ensure that the latest technology in all areas of IT is available to its users. It is one of the total solution providers to the Government and is actively involved in most of the IT enabled applications and has changed the mindset of the working community in the Government to make use of the latest state of the art technology in their day to day activities to provide better services to the citizens. Any service should be accessible by anybody from anywhere at anytime. Even if Internet population is exponentially growing in India, still there is a significant portion of the

people who may not be able to access services for various reasons like limited access to ICT technologies and devices, low literacy, or phobia for Computer etc. Therefore, universal access is still a mirage.

3 Grid Computing: Overview and Suitability for Distributed Information Management Services

Enterprise grid computing is an emerging IT architecture that delivers flexible enterprise information systems that are more resilient and less expensive than traditional legacy systems. In grid computing, groups of independent hardware and software components are pooled and provisioned on demand to meet changing business needs. The accelerating adoption of grid technology is in direct response to the challenges facing information technology (IT) organizations. With today's rapidly changing and unpredictable business climate, IT departments are under increasing pressure to manage costs, increase operational agility, and meet IT service-level agreements (SLAs) [10].

This paper provides an overview of grid computing, highlights the benefits of using it, and describes key grid computing techniques that enable IT resource consolidation, agile IT operations, predictable high performance and scalability, and continuous availability. One way to think about grid computing is as the virtualization and pooling of IT resources compute power, storage, and network capacity, and so on into a single set of shared services that can be provisioned or distributed, and then redistributed as needed. Just as an electric utility uses a grid to deal with wide variations in power demands without affecting customer service levels, grid computing provides IT resources with levels of control and adaptability that are transparent to end users, but that let IT professionals respond quickly to changing computing workloads.

The term utility computing is often used to describe the metered (or pay per use) IT services enabled by grid computing [11]. Cloud computing (where dynamically scalable and often virtualized resources are provided as a service over the internet) is another term that describes how enterprises are using computing resources on both private and public networks over the internet. Because grid computing provides superior flexibility, it is the natural architectural foundation for both utility and cloud computing. As workloads fluctuate during the course of a month, week, or even through a single day, the grid computing infrastructure analyzes the demand for resources in real time and adjusts the supply accordingly.

4 Proposed Grid Computing Based Future NIC

The existing NIC infrastructure in India is primarily provided in a hierarchical manner, where the services are provided either at the grass-root level, in the form of district level information, or at a higher level, like the State-level repositories, ultimately culminating in the national level repository services dealing with pertinent information. Accordingly, the relevant databases are maintained in appropriate databases. But the foremost confrontation is faced whenever the data required is from

a collection of these different databases. Grid computing is a special technology that offers seamless integration to data at different repositories in a seamless, uniform manner, giving the illusion of a single, global data repository. The organizational architecture of such a grid computing based futuristic NIC infrastructure has been shown in figure 3.

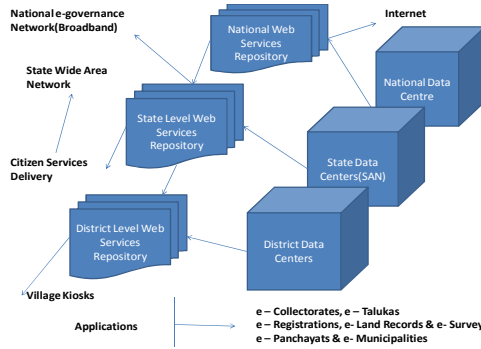


Fig. 3. Proposed E-Governance Grid Web Services Repositories of India

Moreover, the data present at various repositories can be logically aggregated as required to provide higher level web services, like mailing services, network management, and knowledge management and so forth as is presented in figure 4. By means of existing grid middlewares, like the Globus toolkit (GT 4) [12], gridgain [13], etc. such Grid computing based web services repositories can easily be deployed without much of an overhead delivering seamless high level services, like the ones shown in figure 4.

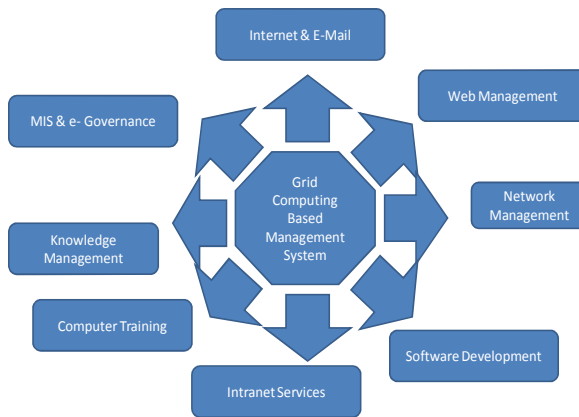


Fig. 4. Grid Computing Based E-governance Management System

5 Conclusion

Changes in this world are obvious and they are difficult to accept, and this is a universal truth. But changes do occur for the betterment with the progress of civilization. India has embraced e-governance and the advantages are visibly perceptible. In course of this paper, it has been highlighted that though here has been major changes in the way that India has been maintaining the information through several ways and means, the national Informatics centers (NICs) being one of the prime sources; yet the information available are segregated like separated islands. In order that India takes full advantage of the available information resources maintained through the NIC infrastructures, it is high time that the latest technologies in ICT horizon be employed. This paper presents the suitability of Grid Computing as an enabling technology in order to achieve futuristic e-governance and management systems. The paper proposes a framework for Grid Web Services Repositories to enhance capabilities of E-Governance in India. It also outlines the salient features of such a futuristic Grid Computing Based E-governance Management System. There are many challenging issues that lie ahead.

References

1. Mahapatra, R., Perumal, S.: e-governance in India: a strategic framework. *International Journal for Infonomics: Special Issue on Measuring e-Business for Development* (January 2006)
2. Signore, O., Chesi, F., Pallotti, M.: E-Government: challenges and opportunities. In: *CMG Italy - XIX Annual Conference* (2005)
3. Monga, A.: E-government in India: Opportunities and challenges. *JOAAG* 3(2) (2008)
4. Bhatnagar, S.: e-government from vision to implementation. Sage publications, New Delhi (2004)
5. Government of India, Information Technology Action Plan: IT for All Indians (2008), <http://it-taskforce.nic.in>
6. Shah, M.: E-Governance in India: Dream or reality? *International Journal of Education and Development using Information and Communication Technology (IJEDICT)* 3(2), 125–137 (2007)
7. Gupta, M.P.: *Towards E-Government Management Challenges*. Tata McGraw-Hill Publishing Company Limited, New Delhi (2004)
8. Kaushik, P.D.: E-Governance: Government Initiatives in India. In: Debroy, B. (ed.) *Agenda for improving Governance*. Academic Foundation in Association with Rajiv Gandhi Institute for Contemporary Studies, New Delhi (2004)
9. Kochhar, S., Dhanjal, G.: E-government Report Card, *Yojna*, New Delhi, vol. 49 (August 2005)
10. Tierney, B., Johnston, W., Lee, J., Thompson, M.: A data intensive distributed Computing architecture for “grid. applications”. *Future Generation Computer Systems* 16(5) (2000)
11. Foster, I., Kesselman, C.: *The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufmann, San Francisco (1999)
12. Globus Toolkit 4.0, <http://www.globus.org/toolkit/> (last accessed on September 23, 2011)
13. GridGain, <http://www.gridgain.com/> (last accessed on September 25, 2011)

A Distortion Free Relational Database Watermarking Using Patch Work Method

R. Arun, K. Praveen, Divya Chandra Bose, and Hiran V. Nath

Amrita VishwaVidhyapeetham, Ettimadai, Coimbatore, Tamil Nadu
{csarunr,praveen.cys,divyacbose,hiranvnath}@gmail.com

Abstract. Database relations are widely used over the Internet. Since these data can be easily tampered with, it is critical to ensure the integrity of these data. In this paper, we propose to make use of fragile watermarks to detect malicious alterations made to a database relation. The proposed scheme is distortion free, unlike other watermarking schemes which inevitably introduce distortions to the cover data. In our algorithm, the watermark is calculated from the linear feedback shift register generating values of the key. Watermarks are embedded and verified in database independently and hence any modifications can be detected.

Keywords: Fragile watermarking, linear feedback shift register, database security, integrity.

1 Introduction

The recent surge in the growth of the Internet results in offering of a wide range of web-based services, such as database as a service, digital repositories and libraries, e-commerce, online decision support system etc. These applications make the digital assets, such as digital images, video, audio, database content etc, easily accessible by ordinary people around the world for sharing, purchasing, distributing, or many other purposes. As a result of this, such digital products are facing serious challenges like piracy, illegal redistribution, ownership claiming, forgery, theft etc. Digital watermarking technology is an effective solution to meet such challenges. A watermark is considered to be some kind of information that is embedded into underlying data for tamper detection, localization, ownership proof, traitor tracing etc. Watermarking techniques apply to various types of host content. Here, we concentrate on relational databases.

Embedding watermarks in database relations is a challenging problem because there is little redundancy present in a database relation. One important property of digital watermarks is invisibility. Usually, in a watermarking scheme, a watermark is embedded by slightly modifying the cover data. This requires that the cover data can tolerate these modifications. In the context of multimedia data, this requirement is not a problem. Since multimedia data are highly correlated, there is a lot of redundant information present in multimedia data [1]. Although compression techniques can remove some of the redundant information, currently, no compression technique is perfect enough to remove all the redundant information. So we move for watermark

embedding. A watermark can be embedded as a part of the redundant information without affecting the quality of the multimedia data. A tuple can be added, deleted, or modified without affecting other tuples. All tuples and all attributes are equally important. There is little redundancy present in the tuples. Thus, it is a challenge to embed an invisible watermark in a database relation.

In general, the database watermarking techniques consist of two phases: watermark embedding and watermark verification. During the embedding phase, a private key K (known only to the owner) is used to embed the watermark W into the original database. The watermarked database is then made publicly available. To verify the ownership of a suspicious database, the verification process is performed where the suspicious database is taken as input, and by using the key K (the same which is used during the embedding phase) the embedded watermark (if present) is extracted and compared with the original watermark information. A suspicious database can be any watermarked database or innocent database, or a mixture of them under various database attacks.

In this paper, we propose a fragile watermarking scheme for detecting malicious alterations made to a database relation. Unlike other watermarking schemes which inevitably introduce distortions to the cover data, the proposed scheme is distortion free.

The rest of this paper is organized as follows. Section 2 gives an overview of the related work. Section 3 explains in detail our proposed fragile watermarking scheme, including watermark embedding and watermark detection. Security analysis of the scheme is provided in section 4. Section 5 concludes this paper with summaries and suggestions for future work.

2 Related Work

In order to achieve the purpose of fragile watermark, authors in [4], [9] proposed watermarking schemes which are able to detect any modifications made to a database relation. These schemes are designed for categorical data that cannot tolerate distortion, hence, the watermark embedding is distortion free. In [4], partitioning of tuples is based on the hash value parameterized with primary key and secret key, whereas in [9], partitioning is based on categorical attribute values. In [4], a watermark of length equal to the number of tuple pairs in the group, is extracted from the group level hash value and for each tuple pair, the order of the two tuples are changed or unchanged according to their tuple hash values and the corresponding watermark bit. Moreover, Li [10] suggests to perform the exchange of tuples' positions based on Myrvold and Ruskeys linear permutation unranking algorithm [8] to increase the embedding capacity

The scheme proposed by [6] aims at maintaining the integrity of the information in the database and is based on public authentication mechanism. The public watermarking scheme by Li and Deng [10] is applicable for marking any type of data including integer numeric, real numeric, character, and Boolean, without fear of any error constraints. The interesting features of this scheme are that it does not use any secret key and can be verified publicly as many times as necessary. The unique watermark key, used in both creation and verification phase, is public and obtained by

one-way hashing. In the algorithm, a cryptographic pseudorandom sequence generator (e.g., Linear Feedback Shift Register) to randomize the order of the attributes and the MSBs of the attribute values are used for generating the watermark W . Any modification to these MSBs introduces intolerable errors to the underlying data and can easily be captured during verification phase. However, alteration of other bits in the data cannot be detected by this scheme.

3 Algorithms

3.1 Design Criteria

This scheme is a fragile watermarking scheme for tamper detection. In this kind of scheme, an attacker will try her best to make modifications to a database relation while keeping the embedded watermarks untouched. The attack is successful if the database relation is modified while the embedded watermarks are still detectable. Thus, in our scheme, the embedded watermarks are designed to be fragile so as to detect any modifications made to a database relation. Given appropriate key and watermark information, a watermark detection process can be applied to any suspicious database so as to determine whether or not a legitimate watermark can be detected.

Table 1. Notations and parameters

Notations	Parameters
γ	number of attributes in the relation
ω	number of tuples in the relation
h_i	tuple hash of the i^{th} tuple in the table or in a group
K	watermark embedding key
W	watermark embedded in the database
W'	watermark extracted from the database
r_i	the i^{th} tuple in the table
A_i	Attribute of the table

3.2 Method 1

Watermark Embedding. Suppose there is a database relation which has a primary key P and γ attributes, denoted by $T(P, A_1, A_2, \dots, A_\gamma)$. Algorithms `Watermark_Generate` describe the watermark generation algorithm and Algorithm `Watermark_Embedding` describes the embedding algorithm. A watermark, the length of which is equal to the number of tuple pairs in the database, is extracted by calculating the bits generated by a LFSR with initial fill as key used for hash calculation. That is, some selected bits from the bits generated by LFSR are put together to form a watermark. Before embedding, a secure tuple hash is computed for each tuple based on the embedding key and all attributes of the tuple. Since the attributes can be of any type, we encode each attribute to an integer. To embed the watermark, for each tuple pair, the order of

the two tuples are changed or unchanged according to the watermark bit to embed. As we can see, since only the order of tuples is changed, the watermark embedded is distortion free. The figure 1 shows an original and its watermarked database which has 8 tuples. For simplicity, we only show hypothetical hash value, not the real value, for each tuple. To embed a watermark we are considering one tuple pair at a time and interchange the tuple if the watermark to embed is 1 and no change if the watermark to embed is 0. In the example if we want to embed 1, then interchange tuple 1 and tuple 2. Then 0 is going to embed, so no interchange in tuple is needed and so on.

Algorithm. Watermark_Generate (K, w)

// K be the key used for hash calculation, w be the number of tuples in the database relation

```

1: knew = LFSR (K)
2: if length (knew) >= w/2 then W = concatenate first w/2 bits of knew
3: else
4:   m=w/2 - length (knew)
5:   W=concatenate of knew and first m bits of knew to make length w/2
6: end if

```

Algorithm. Watermark_Embedding

// K is the key used for hash calculation, w is the number of tuples in the database relation

```

1: for i = 1 to w do
2:   h[j] = HASH (K, ri.A1, ri.A2 ...ri.Aγ)
3: end for
4: sort tuples in database in ascendant order according to their hash
5: W = Watermark_Generate (K, w)
6: for i = 1, j = 1; i <= w; i=i+2, j=j+1 do
7:   if W[j] == 1 then switch the position of tuple i and i+1
8:   else no switching of tuple is needed
9: end if
10: end for

```

Watermark Verification. Algorithms Watermark_Verification describe the watermark detection algorithm. To verify the integrity of a database relation, we need to know K and w. As in watermark embedding, the tuple hash is computed for each tuple. Like watermark embedding, the sorting is a virtual operation and does not involve order change of any tuples. Based on tuple values and the secret embedding key, a hash value is computed and a watermark W' is extracted by comparing the hash values of a pair. If the hash value of first tuple in a pair is greater than second tuple then watermark extracted as 1, otherwise as 0. W is the watermark that is supposed to be embedded. After W' is extracted, it is checked against W. If the two matches, the tuples in the table is authentic; otherwise, it is not.

Algorithm. Watermark_Verification (K, w)

// K is the key used for hash calculation, w is the number of tuples in the database relation

```

1: wt = Watermark_Generate (K, w)
2: for i = 1, j = 1; i <= w; i=i+2, j=j+1 do
3: h[i] = HASH (K, ri.A1, ri.A2 ...ri.Aγ)
4: h[i+1] = HASH (K, ri+1.A1, ri+1.A2 ...ri+1.Aγ)
5: if h[i] <= h[i+1] then W' [j] = 0
6: else W' [j] = 1
7: end if
8: end for
9: if W[j] == W' [j] then return TRUE
10: else return FALSE
11: end if

```

Advantages and Disadvantages

- It will easily detect the tuple addition and deletion easily by checking the length of the database formed or length of the extracted watermark.
- Attribute modification can be detected but with the probability of half.

3.3 Method 2

Watermark embedding. We propose a new method which uses the patch work method technique in watermarking which overcomes the disadvantage of the above method. Here the sorted database is divided into two group based on LFSR states and by using the patch work method technique we find a value d which will be helpful for finding the attribute modification. Then the watermark is embedded into the sorted database. For embedding the watermark the tuples are interchanged in a circular form depending on the watermark to embed which is same as the above method. The group division and the selected patterns are shown in figure 1.

Algorithm. Group_Division(w)

//w is the number of tuples

```

1: select a primitive polynomial to generate maximum length sequence of LFSR (w)
2: get the states of the LFSR (w)
3: b[i] = equivalent decimal representation of states
4: g[j] = select only b[i] which is in between [1, w]
5: return g

```

Algorithm. Watermark_Generate (K, w)

// K be the key used for hash calculation, w be the number of tuples in the database relation

```

1: knew = LFSR (K)
2: if length (knew) >= w+1 then W = concatenate first w+1 bits of knew
3: else m=w+1 - length (knew)
4: W =concatenate of knew and first m bits of knew to make length w+1
5: end if

```

Algorithm. Watermark_Embedding

```
// K is the key used for hash calculation, w is the number of tuples in the database relation
1: g = Group Division (w)
2: for i = 1 to w do
3: h[i] = HASH (K, ri.A1, ri.A2 ...ri.Aγ)
4: end for
5: sort tuples in database in ascendant order according to their hash
//patch work method
6: Take odd number tuple from g as group A
7: Take even number tuple from g as group B
8: Select any random number r
9: for each element in group A, find h' (A) = h (A) + r
10: for each element in group B, find h' (B) = h (B) - r
11: find d = ∑ [ h' (A) - h' (B)]
12: W = Watermark Generate (K, w)
13: order database
14: for i = 1;i<= w+1;i=i+1 do
15: if W[i] == 1 then switch the position of tuple i and i-1 mod w
16: else no switching of tuple is needed
17: end if
18: end for
```

w = 10, Equivalent Binary = 1010 4 bits, Max periodicity = 15 Primitive polynomial = x^4+x+1					
1010	10	0100	4	1110	14
1101	13	0010	2	1111	15
0110	6	0001	1	0111	7
0011	3	1000	8	1011	11
1001	9	1100	12	0101	5
Selected Pattern={10,6,3,9,4,2,1,8,7,5} Group A = {10, 3, 4, 1, 7} Group B = {6, 9, 2, 8, 5} r = 5 $h'(A) = h(A) + r$ $h'(B) = h(B) - r$ $d = \sum h'(A) - \sum h'(B) = -2216$					

Fig. 1. Group Division

Watermark Verification. Algorithms Watermark_Verification describe the watermark detection algorithm. To verify the integrity of a database relation, we need to know K, w, r and d values. First as in the watermark embedding, the tuple hash is computed for each tuple and sorts the table based on the hash values. Then depending on the watermark the tuples will be interchanged back. By applying group division algorithm the new hash value will be found for each group and finally the d value will be find, and check whether it matches with the received one. If it matches then attribute modification is not there. Similarly the tuple addition or deletion is also detected.

Algorithm. Watermark_Verification (K, w, r, d)

```

// K is the key used for hash calculation, w is the number of tuples in the database
relation
// r, d used for patch work method technique
1: for i = 1 to w do
2: h[i] = HASH (K, ri.A1, ri.A2 ...ri.Aγ)
3: end for
4: W = Watermark Generate (K, w)
5: for i = W+1;i>= 1; i=i-1 do
6: if W[i] == 0 then no interchange between i and i-1 tuple
7: else interchange between i and i-1 tuple
8: end if
9: end for
10: g = Group Division (w)
11: Sort the database based on the ascending order of hash values and r be the random
number from embedding side
//patch work method
12: Take odd number tuple as group A and even number tuple as group B from g
13: for each element in group A, find  $h'(A) = h(A) + r$ 
14: for each element in group B, find  $h'(B) = h(B) - r$ 
15: find  $d' = \sum [h'(A) - h'(B)]$ 
16: if  $d' == d$  then no attribute modification
17: else tamper detected

```

Advantages and Disadvantages

- It will easily detect the tuple addition and deletion easily by checking the length of the database formed or length of the extracted watermark
- The Attribute modification can be detected by the introduction of patch work technique

4 Conclusion

In this paper we proposed three different methods for distortion free watermarking techniques for relational databases. The finally proposed method is the one which detects almost all tuple modifications and attribute modifications. Most of the distortion-free watermarking techniques mostly are fragile and aim at maintaining integrity of the database information. Finally, we observe that the usability of the watermarked database and queries still remains an open issue for future research.

References

- [1] Agrawal, R., Kiernan, J.: Watermark relational databases. In: Proc. of the 28th Inter. Conf. On Very Large Data Bases (2002)
- [2] Li, Y., Swarup, V., Jajodia, S.: A robust watermarking scheme for relational data. In: Proc. of the 13th Workshop on Information Technology and Engineering, pp. 195–200 (December 2003)

- [3] Sion, R., Atallah, M., Prabhakar, S.: Rights protection for relational data. In: Proceedings of ACM SIGMOD 2003 (2003)
- [4] Li, Y., Guo, H., Jajodia, S.: Tamper detection and localization for categorical data using fragile watermarks. In: Proceedings of the 4th ACM Workshop on Digital Rights Management (DRM 2004), pp. 73–82. ACM Press, Washington, DC (2004)
- [5] Lin, E., Delp, E.: A review of fragile image watermarks. In: Proc. Of the Multimedia and Security Workshop (ACM Multimedia 1999), October 30–November 5 (1999)
- [6] Sion, R.: Proving ownership over categorical data. In: Proceedings of ICDE 2004 (2004)
- [7] Sion, R., Atallah, M., Prabhakar, S.: Rights protection for relational data. In: Proceedings of ACM SIGMOD 2003 (2003)
- [8] Myrvold, Ruskey: Ranking and unranking permutations in linear time. *Inf. Process. Lett* (2001)
- [9] Bhattacharya, S., Cortesi, A.: A distortion free watermark framework for relational databases. In: Proceedings of the 4th International Conference on Software and Data Technologies (ICSOF 2009), pp. 229–234. INSTICC Press, Sofia (2009a)
- [10] Li, Y., Deng, R.H.: Publicly verifiable ownership protection for relational databases. In: Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security (ASIACCS 2006), pp. 78–89. ACM Press, Taipei (2006)
- [11] Tsai, M., Tseng, H., Lai, C.: A database watermarking technique for temper detection. In: Proceedings of the 2006 Joint Conference on Information Sciences (JCIS 2006). Atlantis Press, Kaohsiung (2006)

A High-Speed Two Dimensional Hierarchical Clustering of Microarray Gene Expression Data

R. Priscilla¹ and S. Swamynathan²

¹ Department of Computer Science and Engineering,
Anna University, Chennai, India
rpriscillaphd@gmail.com

² Department of Information Science and Technology,
Anna University, Chennai, India

Abstract. DNA micro array technology has become the most extensively used functional genomics approach in the bioinformatics field after genome sequencing. Revealing the patterns concealed in gene expression data offers a fabulous opportunity for an enhanced understanding of functional genomics. However, the large number of genes and the difficulty of biological networks greatly increase the challenges of comprehending and interpreting the resulting mass of data, which often consists of millions of measurements. The first step to address this challenge is the use of clustering techniques. Many clustering methods have been devised and used in the analysis of micro array data but less effort has gone into algorithmic speed up of those methods. In this research, quad tree based high-speed two dimensional hierarchical clustering is presented. In the hierarchical clustering process, the construction of the closest pair data structure in each level is the important time factor which determines the processing time of clustering. The proposed high-speed two dimensional clustering process uses the quad tree based data structure for finding the closest pair elements and thus reduces the processing time effectively and produces the better analysis of gene expression data.

Keywords: Clustering, Hierarchical Clustering, Supervised clustering, Overlapping Clustering, Quad Tree.

1 Introduction

Microarrays based global gene expression profiling using is developing as a vital technology for comprehending basic biology of gene function, development, and for finding out new classes of diseases like cancer and for interpreting their molecular pharmacology [1]. Instead of providing full information about genes, microarrays indirectly represent genes through their expressions. [7].

Microarrays have emerged as the standard for simultaneous evaluation of the expression levels of thousands of genes [2]. Clustering techniques play a significant role in discovering sets of objects with identical functions from huge quantities of data [3]. Grouping genes with identical biological functions or categorizing samples with identical gene expression profiles is the usual objective of performing clustering

on microarray gene expression data [4]. Several clustering algorithms have been used for the analysis of microarray gene expression data [5]. Most of the clustering algorithms available these days are distance-based; some examples are Hierarchical clustering, K-means clustering and Self Organizing Map [6] [9].

Hierarchical clustering method extensively used in microarray data analysis combines all data points into a single set by keeping on combining pairs of data points or sets of points adjacent to each other in the feature space [8]. Obtaining the best clustering that signifies a set of patterns in the background of a given distance metric is the objective of hierarchical clustering. This method is commonly based on a similarity or distance measure of the data, like correlation, Euclidean, squared Euclidean, or city-block (Manhattan) distance [13].

From the review given in section 2, it is obvious that data is partitioned by clustering algorithms used in the previous research in such a way that each gene belonged to only one cluster. Some examples for the clustering algorithms that create only one cluster for a gene are K-means algorithm, Hierarchical clustering algorithm, biclustering algorithm, fuzzy k-means algorithm and SOM used in gene expression data. But, these methods have disadvantages when working with microarray gene expression data that gives rise to biological complexity. The nature of proteins and their interactions is the major reason for this. The genes that generate proteins are expected to express with more than one group of genes because proteins generally perform diverse biological roles by interacting with diverse groups of proteins. This explains the inclusion of a gene in more than one cluster of microarray gene expression data. In this research a high-speed two dimensional hierarchical clustering is proposed to represent the existence of genes in one or more clusters consistent with the nature of the gene and its attributes, and prevent biological complexities. The proposed technique uses quad tree based data structure for finding the closest pair and reduces the processing time.

The structure of the paper is organized as follows: A brief review of the researches related to the hierarchical clustering is given in Section 2. The proposed technique for high-speed two dimensional hierarchical clustering is given in Section 3. The experimental results of the proposed approach are presented in Section 4. Finally, the conclusions are given in Section 5.

2 Review of Research

A handful of researches have been presented for clustering micro array gene expression data. Seo Young *et al.* [10] have discussed a broad range of problems like categorization of disease subtypes and tumors in biological and medical research. They have discovered that normalization and extent of noise and clearness for datasets affect the clustering methods that are normally used in micro array data analysis. Carla Layana *et al.* [11] have discussed that simultaneous measurement of the expression levels of thousands of mRNAs has been enabled by micro array technique. Iris Eisenberg *et al.* [12] have discussed that hereditary inclusion body myopathy (HIBM) has been signified by adult beginning gradually developing distal

and proximal myopathy. D'Souza et al. [13] have discussed that comprehending the processes that influence the regulatory networks and pathways controlled inter-cellular and intra-cellular activities has been the objective of Gene Expression Analysis. Liping Jing *et al.* [14] have presented a stable gene selection and efficient cancer prediction.

3 A High-Speed Two Dimensional Hierarchical Clustering

In this research, a high-speed, novel, semi supervised two dimensional quad tree based hierarchical clustering technique has been used for analyzing the gene expression data which includes self clustering of each gene type of clustering elements in vertical dimension and hierarchical clustering of gene type in horizontal dimension. Repeatedly the set of clustering elements are selected randomly from the micro array gene expression database using the index matrix and are clustered using the two dimensional clustering technique. From the resultant clusters, the best 'k' clusters are found out using fitness evaluation. Subsequently, the closest index of all best 'k' clusters are calculated and used to fetch the next set of clustering elements from the database. This process is repeated 'r' times until the optimum cluster is found out.

Let D_{MN} be a database that contains 'M' gene representation of 'N' clustering elements and $d = \{d_{ij} \mid d_{ij} \in D\} \mid 1 < i \leq n; 1 < j \leq N$ be the 'n' type gene representation of 'N' elements, selected randomly from the database D_{MN} using the index $I = \{i_{ij} \mid 1 < M \forall i, j\} \mid 1 < i \leq 1; 1 < j \leq k$. Each value in the 'I', which represents the row index value of Database D_{MN} , must be unique and less than the maximum number of gene representations 'M' in the database D_{MN} .

3.1 Quad Tree Based Two Dimensional Clustering

Many clustering methods have been devised and used in these applications but less effort has expended into algorithmic speed up of these methods. In the hierarchical clustering process, the construction of the closest pair data structure in each level is the important time factor which determines the processing time of clustering. In this research, a quad tree [19] based technique is used to speed up the two dimensional hierarchical clustering. The quad tree based hierarchical clustering is constructed consecutively by inserting the clustering elements one by one into the appropriate node based on the distance. Initially, the entire inner gene values in the every gene type are clustered into four clusters based on quad tree and subsequently every gene types are clustered horizontally for analysis.

3.1.1 Gene Clustering-Vertical Dimension

The basic algorithm for quad tree based inner gene clustering techniques is as follows.

Algorithm. Quad Tree based Inner Gene Clustering

Input: N clustering elements

Step 1: Create root node with empty.

Step 2: Start the first level of hierarchical clustering by inserting ‘n=4’ clustering elements in the root node.

Step 3: Select a clustering element next to be inserted and find the corresponding closest pair element ‘CP_k’ in the level k of the tree by means of Euclidean distance.

Step 4: If the weight of the closest node is less than four, then insert the clustering element as sub node of closest pair element. Else the weight of the closest pair node ‘CP_k’ in the level ‘k’ is greater than four, then find the closest pair node CP_k in the subsequent levels and insert the clustering element as sub node of closest pair element.

Step 5: Repeat steps (3) and (4) until all the elements are clustered.

Output: The resultant clusters.

In the vertical dimension quad tree based clustering; the inner gene values in the every gene type are grouped arbitrarily into four clusters.

3.1.2 Gene Clustering-Horizontal Dimension

Every gene values are clustered inner wise into four clusters according to their distance and then these clusters are get clustered in the horizontal dimension for analyzing the gene values. The distance is calculated as follows.

$$Ed(p, q) = \sqrt{\sum (p - q)^2}$$

Where p is the clustering elements in ‘d’ and q is the clustered element in the kth level of the quad tree. The distance between two elements P and Q is how much is the root of, sum of, square of, deviation among P and Q. The first level of horizontal wise hierarchical clustering starts with the selection of two elements having greatest distance and inserted into the root node. Using the Euclidian distance, the closest pair node of next gene representation element in the kth level is find out and inserted into the corresponding node if the weight of the node is less than four. Otherwise, it finds out the closest pair element in the subsequent levels and inserts the node. Finally, every gene types are clustered based on quad tree.

3.2 Fitness Evaluation

Let C be the resultant cluster, the fitness of C is calculated as follows.

$$\frac{1}{0.1 + \sum w(C)}$$

Where $w = \begin{cases} 1 & \text{if } C[i] = R^{def} \\ 0 & \text{otherwise} \end{cases}$ is the weight of the each cluster element and R^{def} is

the defined cluster used for the semi supervised hierarchical clustering. If an element in the resultant cluster is in the defined cluster R^{def}, then the weight of the element

will be 1 and 0 otherwise. The two dimensional clustering process and fitness evaluation are processed for every row of the index 'I' and $K = \{C_i \mid 1 < i \leq l\}$ is the resultant cluster. From the resultant cluster set 'K', the best 'k' clusters having the highest fitness value are chosen and the closest index of all best 'k' clusters are calculated which are then used to fetch the next set of clustering elements from the database. The clustering is processed repeatedly until the optimum cluster is found and using the presence of gene is analyzed.

4 Experimental Results

The proposed technique is implemented in the MATLAB platform (version 7.11) with the system configuration Intel(R) Core(TM) i5 CPU,650@3.20GHz,3.19 GHz, 3.17 GB of RAM and it is evaluated using the micro array gene expression data of human acute leukemia. The two dataset of standard leukemia for training and testing is obtained from [23] and performance of the proposed technique in clustering ground truth data cancer classes, namely, acute myeloid leukemia (AML) and acute lymphoblast leukemia (ALL) are demonstrated. The two training leukemia dataset are partitioned again and turned to four set (dataset_1, dataset_2, dataset_3 and dataset_4) each having N values (20, 18, 18, 17) respectively. This high dimensional training dataset is subjected to clustering for analyzing the presence of micro array genes in more than one cluster.

In this clustering technique, an adaptive approach is followed to dynamically define the number of clusters that must be generated from the micro array gene expression dataset. The proposed technique, which is a multi-stage clustering technique, performs clustering at different levels. The gene values in the every gene type are clustered internally into four clusters and then the every gene types are again clustered horizontally for analysis. In the inner gene clustering the root node is created first and then the four clustering elements which are selected randomly are inserted into the root node. The next element to be inserted is selected and using the Euclidian distance the closest pair node of this element the first level is find out and inserted into the corresponding node if the weight of the node is less than four. If the weight of the node is greater than or equal to four means it finds out the closest pair element in the subsequent levels and insert the node. This process is repeated until all the elements get clustered. Thus every gene representation values are clustered into four clusters. As like the inner gene clustering the every gene representation values are clustered horizontally for analyzing.

4.1 Performance Evaluation

The performance of the proposed two dimensional hierarchical data clustering technique is evaluated on clustering ground truth data of the cancer classes, namely, acute myeloid leukemia (AML) and acute lymphoblast leukemia (ALL) using Precision, Recall and F-measure [16] [17] subsequently these values are compared with the Precision, Recall and F_measure values of Hierarchical clustering without quad tree. We have used the Precision, Recall and F-measure described in [16] [17] for evaluating the performance of the proposed incremental text clustering approach.

Precision and recall values of the clusters obtained by the proposed technique are given in the Table 1 and Fig 1(a) illustrate the corresponding graph. Precision and recall values of the clusters obtained by the Hierarchical Clustering without Quad Tree are given in Table 2 and Fig1(b) illustrates the same.

Table 1. Precision, Recall and F Measure of the Clusters_ Hierarchical Clustering with Quad Tree

Dataset	Cluster	Precision	Recall	F measure
Dataset_1	C1	0.8667	0.9286	0.8966
	C2	0.8000	0.6667	0.7273
Dataset_2	C3	0.9091	0.7692	0.8333
	C4	0.5714	0.8000	0.6667
Dataset_3	C5	0.9091	0.9091	0.9091
	C6	0.8571	0.8571	0.8571
Dataset_4	C7	0.8000	0.8000	0.8000
	C8	0.7143	0.7143	0.7143

Table 2. Precision, Recall and F Measure of the Clusters_ Hierarchical Clustering without Quad Tree

Dataset	Cluster	Precision	Recall	F measure
Dataset_1	C1	0.5000	0.2143	0.3000
	C2	0.2143	0.5000	0.3000
Dataset_2	C3	0.7000	0.5385	0.6087
	C4	0.2500	0.4000	0.3077
Dataset_3	C5	0.8182	0.8182	0.8182
	C6	0.7143	0.7143	0.7143
Dataset_4	C7	0.5000	0.3000	0.3750
	C8	0.3636	0.5714	0.4444

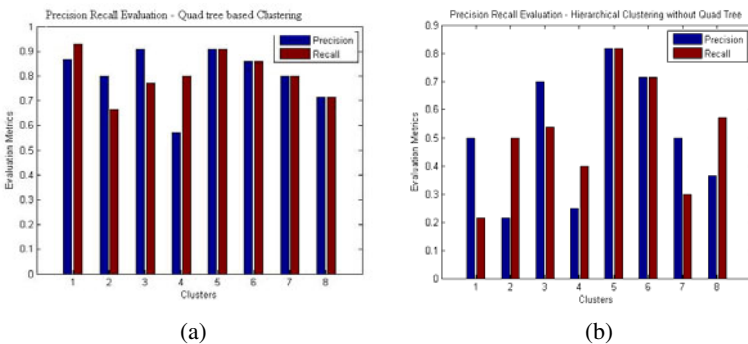


Fig. 1. (a) Precision Recall Evaluation of Quad tree based technique, (b) Precision Recall Evaluation of the Hierarchical Clustering without Quad Tree

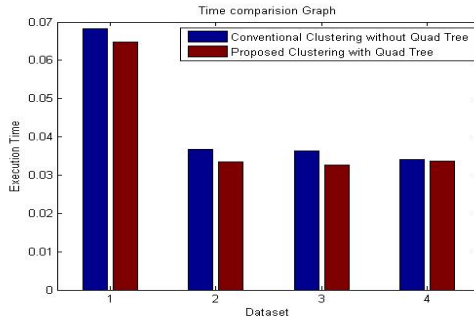


Fig. 2. Time comparison Graph

The performance of the proposed two dimensional hierarchical data clustering with Quad Tree approach is also evaluated by comparing its processing time. From the Fig 2 it is obvious that the hierarchical clustering with Quad tree technique performs clustering faster than the hierarchical clustering without Quad tree.

5 Conclusion

An innovative high-speed two dimensional hierarchical clustering technique to handle the micro array genes that exists in more than one cluster is proposed in this paper. Genes are effectively expressed by the proposed technique which avoids biological complexities. By using the quad tree data structure for finding the closest pair, the proposed system performs faster than the existing system. The clustering performance of the proposed technique was visualized by implementing it in MATLAB. Experimental results on real life datasets have proved that the Hierarchical clustering with Quad tree technique is more effective and faster than the Hierarchical clustering without Quad tree.

Acknowledgements. The Authors expresses their sincere thanks to the Department of Information Science and Technology, Anna University Chennai for providing necessary facility to conduct the research work.

References

1. Liang, J., Kachalo, S.: Computational analysis of microarray gene expression profiles: clustering, classification, and beyond. *Chemometrics and Intelligent Laboratory Systems* 62(2), 199–216 (2002)
2. Cvek, U., Trutschl, M., Stone II, R., Syed, Z., Clifford, J.L., Sabichi, A.L.: Multidimensional Visualization Tools for Analysis of Expression Data. *World Academy of Science, Engineering and Technology* 54(50), 281–289 (2009)
3. Kim, S.Y., Choi, T.M.: Fuzzy Types Clustering for Microarray Data. *World Academy of Science, Engineering and Technology* 4, 12–15 (2005)

4. Wu, X., Chen, Y., Brooks, B.R., Su, Y.A.: The Local Maximum Clustering Method and Its Application in Microarray Gene Expression Data Analysis. *Eurasip Journal on Applied Signal Processing* (1), 53–63 (2004)
5. Chen, G., Jaradat, S.A., Banerjee, N., Tanaka, T.S., Ko, M.S.H., Zhang, M.Q.: Evaluation and Comparison of Clustering Algorithms in Analyzing ES Cell Gene Expression Data. *Statistica Sinica* 12, 241–262 (2002)
6. Qin, Z.: Clustering microarray gene expression data using weighted Chinese restaurant Process. *Bioinformatics* 22(16), 1988–1997 (2006)
7. Grudz, I., Ihnatowicz, S., Siddiqi, A., Akhgar: Mining Genes Relations in Microarray Data Combined with Ontology in Colon Cancer Automated Diagnosis System. *World Academy of Science, Engineering and Technology* 16(26), 140–144 (2006)
8. Wang, R., Scharenbroich, L., Hart, C., Wold, B., Mjolsness, E.: Clustering Analysis of Microarray Gene Expression Data by Splitting Algorithm. *J. Parallel Distrib. Comput.* 63(7-8), 692–706 (2003)
9. Lee, M., Kim, Y.-M., Kim, Y.J., Lee, Y.-K., Yoon, H.: An Ant-based Clustering System for Knowledge Discovery in DNA Chip Analysis Data. *World Academy of Science, Engineering and Technology* 29(48), 261–266 (2007)
10. Kim, S.Y., Hamasaki, T.: Evaluation of Clustering based on Preprocessing in Gene Expression Data. *International Journal of Biological and Life Sciences* 3(1), 48–53 (2007)
11. Layana, C., Diambra, L.: Dynamical Analysis of Circadian Gene Expression. *International Journal of Biological and Life Sciences* 3(2), 101–105 (2007)
12. Eisenberg, I., Novershtern, N., Itzhaki, Z., Becker-Cohen, M., Sadeh, M., Willems, P.H.G.M., Friedman, N., Koopman, W.J.H., Mitrani-Rosenbaum, S.: Mitochondrial processes are impaired in hereditary inclusion body myopathy. *Human Molecular Genetics* 17(23), 3663–3674 (2008)
13. D’Souza, Sekaran, C., Kandasamy: A Phenomic Algorithm for Reconstruction of Gene Networks. *International Journal of Biological and Life Sciences* 4(2), 76–81 (2008)
14. Jing, L., Ng, M.K., Zeng, T.: Novel Hybrid Method for Gene Selection and Cancer Prediction. *World Academy of Science, Engineering and Technology* 62(89), 482–489 (2010)
15. ALL/AML datasets from <http://www.broadinstitute.org/cancer/software/genepattern/datasets/>
16. Larsen, B., Aone, C.: Fast and Effective Text Mining Using Linear-time Document Clustering. In: *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, California, United States, pp. 16–22 (1999)
17. Steinbach, M., Karypis, G., Kumar, V.: A Comparison of Document Clustering Techniques. In: *Proceedings of the KDD-2000 Workshop on Text Mining*, Boston, MA, pp. 109–111 (2000)
18. Chakraborty, A., De, S.K., Dasgupta, R.: Balancing of Quad Tree Using Point Pattern Analysis. *World Academy of Science, Engineering and Technology* (2011)

Achieving Service Level Agreement in Cloud Environment Using Job Prioritization in Hierarchical Scheduling

Rajkumar Rajavel and T. Mala

Department of IST, Anna University,
Chennai, Tamil Nadu, India
rajkumarprt@gmail.com,
malanehru@annauniv.edu

Abstract. One of the challenging issues in Cloud computing Environment is meeting the Service Level Agreement (SLA). The SLA is an agreement signed between the service provider and the service consumer for accessing the service provided by the service provider over the internet. We can investigate the negotiation strategy between the service provider and the service consumer through the third party called a Broker. In many approaches SLA is designed and trusted through the measurement of various non-functional requirements such as response time of job, CPU usage, memory usage and the storage used by the consumer. Main focus of the business process is satisfying the customer need by quick response. In our proposed approach for satisfying the service consumer (customer), parameter such as response time of deadline based job is considered. The response time of the job is affected due to improper scheduling of the job. Therefore a novel hierarchical scheduling with job prioritization is used to give more priority for deadline based jobs. This approach will satisfy the service consumer and meet the SLA by increasing the performance of the scheduling algorithm.

Keywords: Service Level Agreement, service provider, service consumer, negotiation, response time, hierarchical scheduling, job prioritization.

1 Introduction

Cloud computing is a general term for “anything” that can be accessed as the service over the internet. The services can be Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS) and Storage as a service [10]. Cloud computing is a new technology which initiate the new way or new trend of computing where the readily available resources are accessed as a service over the internet. The main feature of this cloud computing is “pay-for-use” technique where the service consumer can pay the amount in the online for the number of resource instantly used and the duration of time the resources are accessed. The key properties of the cloud computing is user centric, task centric, powerful, accessible, intelligent and programmable. The concept of cloud computing is the next step to group collaboration. In-order to do this type of project, we need to deploy or house the project to the cloud so that the project can be accessed from anywhere from the internet enabled locations.

The main advantages of the cloud computing are listed as lower cost computer for use, improved performance, lower IT infrastructure cost, lower software cost, instant software updates, fewer maintenance issue, increased computing power, unlimited storage capacity, increased data safety and improved compatibility between operating system. There are certain limitations in cloud environment such as it requires constant internet connection, availability of service, scalable storage, software licensing and etc, which leads to various research activities [6], [7]. In our proposed approach the scenario considered is software licensing which means pay-for-use license. It is an agreement signed between the service provider and the service consumer as a software document with the negotiation between them.

The main purpose of the SLA licensed software document in business model is to ensure the guarantee of both the service consumer and service provider on accessibility of resource and completion of job respectively. Here the service provider has to complete the job (task or request) assigned by the service consumer within the stipulated time as mentioned in the SLA [1], [3]. Similarly the service consumer has to utilize the resource facility with proper CPU speed, Cache memory, number of nodes or Virtual Machine (VM) and the storage as mentioned in the SLA document. Sometimes the service provider may fail to meet the SLA as mentioned in the document because of the unavailability of the resource and overload of the resource due to improper scheduling of jobs. So a novel hierarchical scheduling with multilevel feedback queue is proposed to meet the SLA. The main purpose of the multilevel feedback queue is to preempt the dead line based job for the execution through which we can meet the SLA and can satisfy the consumer. By means of customer satisfaction we can obviously increase the number of cloud user and increase the productivity of the business.

2 Related Work

The important issues in Cloud are performance degradation in the cloud business due to customer dissatisfaction (when the SLA is not meet by the service provider). There are several approaches to meet the SLA between the service provider and the service consumer [5]. In the paper [2], deadline-aware heuristic approach is used which will serve the deadline based job first. It is very difficult to prioritize the deadline based job using a single queue implementation. So in our approach multilevel feedback queue is used for giving more priority for deadline based jobs. The First Come First Served, Shortest Job First and Heuristic Cost Based Scheduling algorithms were implemented and their performances are evaluated [8]. Here the quality of service is measured using the response time of individual jobs. Even if a single job fails to respond within the stipulated time it leads to violation of SLA. Sometime this approach may lead to violation of SLA because it estimate the heuristic cost by using the waiting time and it give more priority to the job which is having less heuristic cost. Main reason here is some jobs might have less heuristic cost and it is not the deadline based job, but this job is given more priority according to this approach. But some deadline based jobs might have more heuristic cost which will suffer by starvation due to less priority over the heuristic cost estimation.

3 Need for Job Prioritization in Hierarchical Scheduling

In the cloud environment to improve the performance, scheduling is done in hierarchical manner by using scheduler in both Cloud Controller (CLC) and Cluster Controller (CC) level. The jobs dispatched by the CLC will be queued in the cluster node. Here the job will be executed one after another in the First Come First Serve (FCFS) basis in the Round Robin (RR) fashion and this may lead deadline based jobs waiting for long time in the queue based on its position in the queue [4]. This situation will obviously leads to SLA violation and thus it dissatisfies the consumer (not meeting the SLA). In the proposed approach hierarchical scheduling is implemented by using job prioritization in both CLC and CC level to avoid SLA violation.

4 Hierarchical Scheduling in the Cloud Environment

An overview of the Cloud Environment Model is shown in Fig1. It consists of service provider, service consumer, cloud resources, SLA document as a negotiation process and a third party for service guarantee. The service provider might be of different people like Amazon EC2, Microsoft Azure, Google App Engine, Google Apps, Salesforce.com and Microsoft Online Services. Some of the services provides by these service providers are IaaS, PaaS, SaaS and Storage as a Service. The service consumers are the end user who can instantly get the huge computational resource by subscribing to the service provider by connecting to the internet. Consumer can instantly get the resource as a service and start running his application. If the application requires more computational power, the resources can be increased on demand by creating the Virtual Machine and also decreases the resource to certain number by destroying the Virtual Machine. Since the cloud resources are elastic in nature either the resource can be increased or decreased based on the needs of the user application running in the cloud environment. SLA document is a contract which specifies a set of application-driven requirement such as contract duration, estimated number of jobs, estimated memory requirement number of resource (VMs).

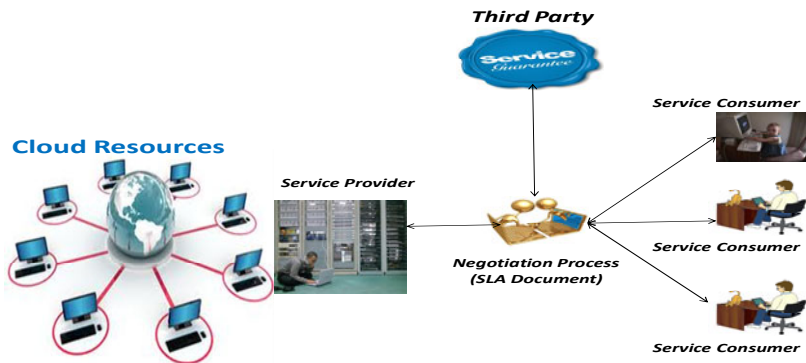


Fig. 1. Overview of Cloud Environment

In this paper Hierarchical Scheduling is proposed in the Cloud Environment as shown in Fig.2 with meta-level scheduling in Cloud Controller (CLC) and local-level scheduling in the Cluster Controller (CC). Service Level Agreement is an agreement signed between the service provider and the service consumer through the negotiation process. If any one of the job is failed to complete within the stipulated time it leads to violation of SLA. We always cannot guarantee the services provided by the service provider and for negotiation with provider we have to depend on the third party. The cloud resources are the group of servers, desktops and PCs which are geographically distributed across the world and connected by the communication media such as wired or wireless. The structure of the cloud resources and its various components such as Cloud Controller, Cluster Controller and Node Controller are organized in the hierarchical structure as shown in Fig. 2.

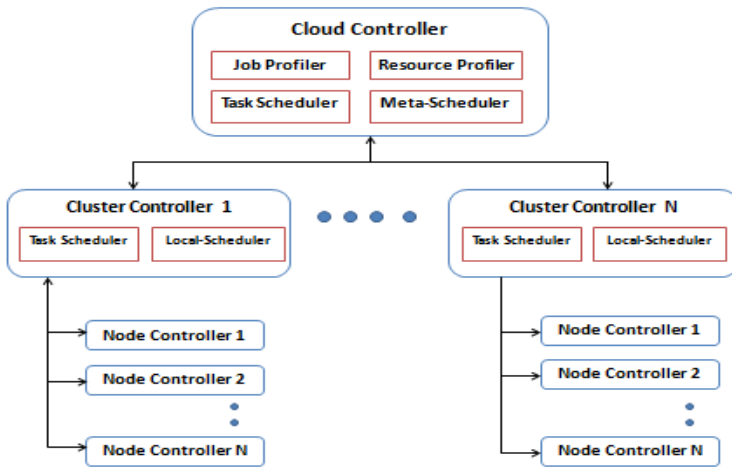


Fig. 2. Structure of Cloud Resources

The main work of the Cloud Controller is to manage and control all the underlying Cluster Controller in the cloud. In addition to that it will schedule all the jobs obtained from the service consumer using the Scheduler component and maintains the result of each job and also update the job status in the Job Profiler component. This Job Profiler will maintain the result and status of all the active jobs which is in execution. It also contains useful information about the non active jobs which are waiting in the queue. The operation of the Cloud Controller will work as shown in Fig. 3. The users request or job will fall into the Cloud Controller Job Queue and for every time interval or instance the Job Puller will pull the job from the Job Queue and sent to the Task Scheduler. Based on the type of job the Task Scheduler will prioritize the job for scheduling process. If the job is deadline or interactive based then it is pushed to high priority queue Q1. In case the job is shortest job it will be pushed to next priority queue Q2, otherwise it will be directly pushed to low priority queue Q3. Since the queue Q1, Q2 and Q3 is connected with the feedback, suppose if one job is moved from the Q1 to meta-scheduler then automatically job from Q2 will be moved to backend of Q1 and similarly the job from Q3 is moved to Q2.

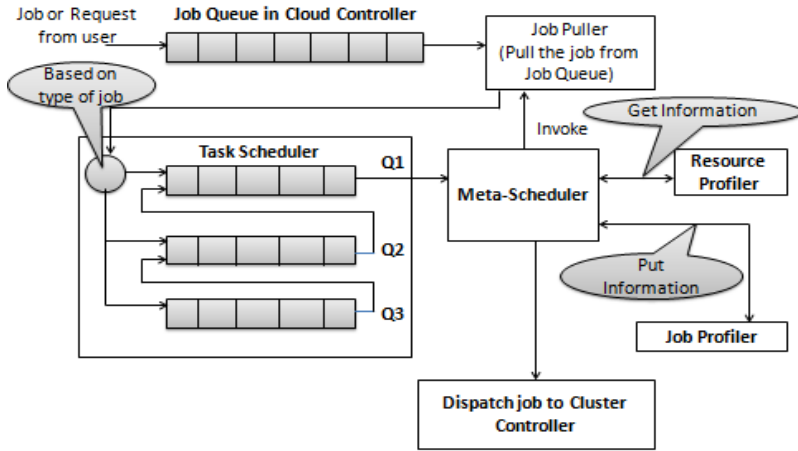


Fig. 3. Operation of Cloud Controller

The Meta-Scheduler will obtain the jobs from Q1 and schedule the job to Cluster Controller by using Load Aware Scheduling Algorithm as shown in Algorithm 1. It will get the resource information from the Resource Profiler and update job information in the Job Profiler during the scheduling process. Suppose if the meta-scheduler find the empty queue in Task scheduler, then it will invoke the Job Puller to pull next set of job from the Job Queue. Finally the job dispatched from the Cloud Controller will fall into the Cluster Controller Job Queue. The Load Aware Scheduling Algorithm in the meta-scheduler will estimate the Load Cost of resource $LC(R)$ by using the equation (1) and (2) as follows,

$$LC(R_i) = QL(R_i) / [\sum NC(R_i)] \quad (1)$$

Where $QL(R_i)$ denotes Queue Load in the resource R_i and it is estimated by using the number of Virtual Machines required by each job as follows,

$$QL(R_i) = \sum [NVM(J_1), NVM(J_2), \dots, NVM(J_n)] \quad (2)$$

Where $NVM(J_1)$ represents the number of VM required by Job1 and 'n' denotes the number of jobs waiting in the resource R_i .

Algorithm 1. Load Aware Scheduling Algorithm

```

Begin
Get job from Q1
for ( each job)
    Identify the job requirement
    Query the resource profiler for updated resource information
    Estimate the Load Cost for all the resource or cluster controller
    Identify the resource having less Load Cost
    Select the matched resource
    Dispatch the job to matched resource
End

```

The operation of the Cluster Controller will work as shown in Fig. 4. The job is pulled from the Job Queue and based on the type of job the Task Scheduler will prioritize the job to Local-Scheduler. Then the Local-Scheduler will follow the FCFS scheduling and finally dispatch the job to corresponding matched Virtual Machine (VM) present in the Node Controller.

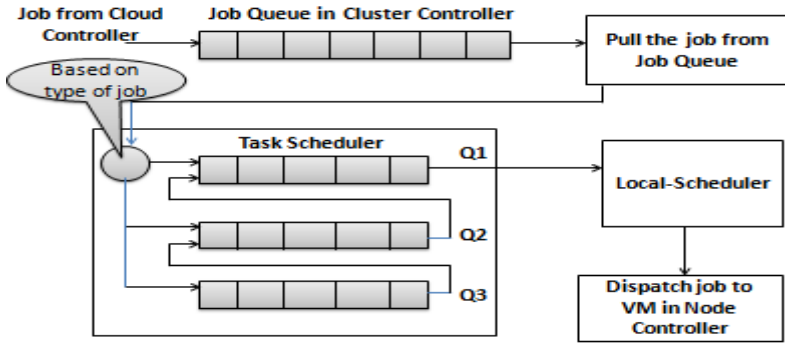


Fig. 4. Operation of Cluster Controller

Since the job is prioritized in hierarchical scheduling, the deadline based jobs will be completed within the stipulated time which leads to meet the SLA made between the provider and the consumer. Job execution and job management in the individual node will be taken care by the Node Controller component. Finally the corresponding output of the job will be dispatched to the end user through the cloud controller component. Based on the response time of the job the third party broker will announce the service guarantee to the job by verifying resource, memory and the estimation cost of the resource used by the user as specified in the SLA document during the negotiation process. If all the jobs are completed within the stipulated time as specified in the SLA document then the result is intimated and the service is guaranteed by the service provider. In case if the job is not completed within the time then the result is specified as SLA violation of service provider. Suppose if the service or resource used by the consumer is more than the specification in the SLA document then it leads to the service consumer or user SLA violation.

5 Experimental Results and Performance Evaluation

In the result phase we have simulated the result using cloudsim toolkit by exploiting five resources and fifty jobs for exactly showing the scheduling algorithm with its response time. Job information present in the job profiler is listed as shown in the Table1. From the table it is clear that only five jobs are deadline based jobs and remaining forty five jobs are non deadline based jobs. All the fifty jobs are submitted to the five available resources only. If you use the FCFS and SJF algorithm all the job requirement cannot be meet, since the deadline based jobs are not given any priority and hence it result in the violation of SLA signed between the service provider and the service consumer [4]. If the same number of job is executed using Load Aware Scheduling Algorithm and Job prioritization in the hierarchical scheduling will result

in the situation where deadline based jobs will be given more priority over the other jobs and it complete the jobs within the stipulated time as specified in the user job requirement. So by using this scheduling algorithm SLA can be achieved easily and satisfy the customer in the cloud business.

Table 1. Job Information

Job ID	Job Size	Completion Time (Deadline of Job)	Execution Time
1 to 10	30 MB	NIL	30 min
11 to 25	20 MB	NIL	30 min
26	30 MB	120 min	30 min
27	20 MB	120 min	40 min
28	30 MB	120 min	50 min
29	20 MB	120 min	40 min
30	30 MB	120 min	40 min
31 to 50	10 MB	NIL	30 min

The results of FCFS, SJF and Load Aware Scheduling (LAS) Algorithm of Hierarchical Scheduling is compared with respect to Completion time of the jobs and the performance measure is also represented as graph as shown in Fig5. In order to show the exact result and its problem definition the results of the deadline based jobs are considered for performance evaluation. From the figure it is clear that only job with ID 26, 27, 28, 29 and 30 are the deadline based jobs.

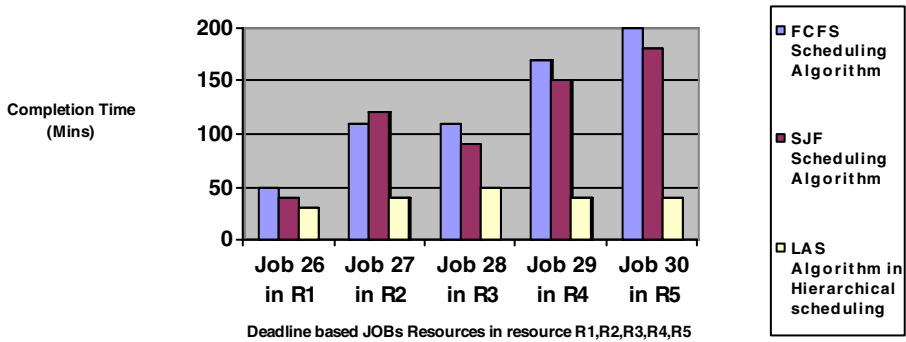


Fig. 5. Performance Evaluation of FCFS, SJF and Multilevel Feedback queue Scheduling Algorithm

The performance of the LAS Algorithm which is proposed in Hierarchical Scheduling works much better than the existing FCFS and SJF Algorithm by completing all the jobs within the stipulated time [9]. Hence from the above performance evaluation it is clear that the proposed Hierarchical Scheduling will increase the performance in the cloud environment and increase the customer satisfaction by quickly responding to the user jobs. By completing all the jobs within the stipulated time SLA is achieved in the cloud.

6 Conclusion and Future Work

In this paper hierarchical scheduling is presented which helps in achieving Service Level Agreement with quick response from the service provider. In our proposed approach Quality of Service metric such as response time is achieved by executing the high priority jobs (deadline based jobs) first by estimating job completion time. Here the priority jobs are spawned from the remaining job with the help of Task Scheduler which increase the performance of the cloud business by quickly responding to the customer. Hence this novel approach provides quick response time comparing to the existing approaches by meeting the SLA in the cloud Environment. In future, multifunctional request handler will be integrated in the cloud controller for handling user jobs in different ways. Here the user can submit the job either through the User Interface provided by the provider or by using the Job Submission Description Language. Thus the multifunctional request handler will help in handling different way of job submission through SOAP request.

References

1. Andrzejak, A., Kondo, D., Yi, S.: Decision Model for Cloud Computing. In: 18th Annual IEEE/ACM International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (2010)
2. Nazir, A., Liu, H., Sorensen, S.-A.: Service Level Agreements in Rental-based Systems. In: 10th IEEE International Conference on Computer and Information Technology (2010)
3. Reig, G., Alonso, J., Guitart, J.: Prediction of Job Resource Requirements for Deadline Schedulers to Manage High-Level SLAs on the Cloud. In: 9th IEEE International Symposium on Network Computing and Applications (2010)
4. Moon, H.J., Chi, Y., Hacigumus, H.: SLA-Aware Profit Optimization in Cloud Services via Resource Scheduling. In: IEEE 6th World Congress on Services (2010)
5. Van Nguyen, H., Tran, F.D., Menaud, J.-M.: SLA-aware Virtual Resource Management for Cloud Infrastructures. In: IEEE 9th International Conference on Computer and Information Technology (2010)
6. Oriol Fito, J., Goiri, I., Guitart, J.: SLA-driven Elastic Cloud Hosting Provider, 18th Euromicro Conference on Parallel, Distributed and Network-based Processing (2010)
7. Hima Prasad, K., Faruque, T.A., Venkata Subraminiam, L., Mohania, M.: Resource Allocation and SLA Determination for Large Data Processing Services Over Cloud, IEEE International Conference on Services Computing (2010)
8. Bloor, K., Chirkova, R., Viniotis, Y.: Dynamic request allocation and scheduling for context aware applications subject to a percentile response time SLA in a distributed Cloud. In: 2nd IEEE International Conference on Cloud Computing Technology and Science (2010)
9. Bloor, K., Chirkova, R., Salo, T., Viniotis, Y.: Heuristic-based request scheduling subject to percentile response time SLA in a distributed cloud. In: IEEE Globecom 2010 Proceedings (2010)
10. Alhamad, M., Dillon, T., Chang, E.: Conceptual SLA Framework for Cloud Computing. In: 4th IEEE International Conference on Digital Ecosystem and Technologies (2010)

Runtime Estimation Aware Scheduling Algorithm for Handling Deadline Based Jobs in Grid Environment

Rajkumar Rajavel and T. Mala

Department of IST, Anna University,
Chennai, Tamil Nadu, India
rajkumarprt@gmail.com,
malanehru@annauniv.edu

Abstract. One of the major issues in Grid Environment is scheduling the deadline based jobs in the meta-scheduler level. In many approaches Bulk Scheduling is done in meta-scheduler by evenly distributing all the jobs present in the request handler queue to the available resources. All the submitted jobs will fall into the resource queue and get executed in the First Come First Served (FCFS) or Shortest Job First (SJF) fashion and it will works better for non-deadline based jobs where prioritization of job is not required. This type of scheduling algorithm will leads to starvation and increases the waiting time of deadline based jobs, and finally results in user dissatisfaction. Moreover, these types of scheduling algorithm in the meta-scheduler will not provide the feasible solution to the deadline based job. So a Runtime Estimation Aware Scheduling is proposed in the meta-scheduler, which provides high priority for deadline based jobs to dispatch first, during the scheduling process. Most algorithm will works good if the job have the runtime input and may not good enough if the job doesn't have the runtime input. Our scheduling algorithm will works well with irrespective of jobs nature (either job with runtime or without runtime), obviously reduces the waiting time of deadline based (high priority) jobs and lights up the grid providers to achieve Service Level Agreement (SLA).

Keywords: Meta-Scheduler, Deadline based jobs, Bulk Scheduling, Runtime Estimation Aware Scheduling, Waiting time and SLA.

1 Introduction

Grid computing is the dynamic and coordinated resource sharing to solve the problem by dynamically linking the various resources of the Virtual Organizations. Here the Virtual Organization represents the group or an individual who engaged in designing rules for the sharing of resource. According to Ian Foster and Kesselman, Grid Computing is hardware and software infrastructure which offers a cheap, distributable, coordinated and reliable access to powerful computational capabilities [4]. Since multiple grid or any applications may require numerous resources which is often not available for them so that in order to allocate resource to input jobs, having a scheduling system in the meta-scheduler is essential. Because of vastness and

separation of resource in the computational grid, meta-scheduling is seems to be most important issues. In the grid environment Bulk Scheduling is mostly often preferred in the Meta-Scheduler level for scheduling the tasks or jobs. Since more number of jobs is arrived to meta-scheduler it cannot be scheduled one after another. So as soon as the job is arrived to the meta-scheduler it should be distributed to the underlying resources as early as possible [10], [2], [13]. This resource will adopt various scheduling algorithms for the execution of the jobs. As per the grid environment is concern each and every individual personal computer and cluster computer is considered as a resource. Cluster computer consist of one head node and any number of computing nodes. Usually job is submitted to the head node of the cluster and from the head node job is distributed over the computing nodes for running the jobs. Here the main role of the head node is to distribute and maintain the job running in the computing nodes. In the cluster nodes, if any problem occurs in computing nodes head node will take the decision to migrate the job to other computing node by using the fault tolerance technique [9], [11], [3].

In the Grid Environment for scheduling the jobs in the meta-scheduler, there might be lots of resource available for the execution of job and hence have an issue of choose the best resource for deadline based jobs. To sort out this issue it is mandatory to know about the nature of grid environment whether the meta-scheduler maintains the jobs history or not. Because, some scheduling strategy like FCFS, JR-backfilling and SLOW-coordinated will works well in the environment were the runtime of the job is known prior to scheduling, either the runtime of job is indicated in job input or the job runtime is estimated from the job history [5]. Similarly the Application Demand Aware algorithm and DIANA Scheduling algorithm works well only in the heuristic environment in which jobs runtime is known for exploiting in estimation of jobs completion time [6], [1]. So novel scheduling algorithms is proposed in the meta-scheduler to adapt in different nature of grid environment. Here the Runtime Estimation Aware Scheduling algorithm is proposed to improve the efficiency of the scheduling strategy in the grid environment where the history of jobs is maintained. Hence the overall performance of the meta-scheduler is improved by the proposed algorithms by reducing the waiting time of the jobs and moreover helps in achieving the Service Level Agreement (SLA) made in the Grid Environment.

2 Related Work

The important issue in grid environment is scheduling deadline based jobs which in turn leads to performance degradation and users dissatisfaction due to improper scheduling of the jobs [1]. In the previous paper, Dynamic Load Balancer algorithm proposal is done for estimating the resource load by using little's formula which will evenly distribute the jobs from the meta-scheduler to the underlying resource. It works well for handling non-deadline based jobs [12]. One more important issue in the scheduling is guaranteed to complete the execution of the jobs within the deadline given by the user for achieving the SLA made between the user and the provider. There are so many scheduling algorithms exist in the meta-scheduler level as shown in paper [13], [8], [6] which provides feasible solution for the non-deadline based jobs

and it is not guaranteed for the completion of jobs within the stipulated time. In paper [7] List Scheduling (LS) algorithm is proposed which gives better in the situation where the meta-scheduler have limited jobs and that too with non deadline based jobs. Main reason for this problem is there may be the possibility to have too many number of jobs waiting in the meta-scheduler where cannot be able to prioritize the deadline based jobs and finally leads to increase in waiting time of job. In paper [3] Proposed Scheduling Algorithm (PSA) is used which will works efficiently in the case were the runtime of the job is known in advance to the user. This algorithm will not work well in the situation where the runtime of the job is not specified in the user's job specification. So in-order to handle this situation runtime estimator is used in meta-scheduler level to estimate the runtime of the job which is not having the execution time or runtime in the job specification.

3 Need of Runtime Estimation Aware Scheduling

The Runtime Estimation Aware Scheduling will helps in making intelligent decision regarding the selection of best resource to meet the user requirement, by estimating completion time of the job using waiting time of job in queue and runtime or execution time of jobs in each resource. This scheduling algorithm in turn exploits the Runtime Estimator component for estimating the runtime of jobs whose runtime is not mentioned in the job specification. These estimations are used by the scheduling algorithm for predicting the completion time of deadline based jobs to choosing the best resource for job submission. By exploiting this Runtime Estimation Aware Scheduling in the meta-scheduler can make the decision making process as good as possible. This runtime estimator is not an actual component of the meta-scheduler, but it is added in our proposed model to improve the efficiency of the meta-scheduler. This scheduling approach works well for the environment were the job has runtime specification and as well as the meta-scheduler were the job history is maintained.

4 Meta-Scheduler in the Grid Environment

In this paper, Runtime Estimation Aware Scheduling and Multilevel Feedback Queue Scheduling algorithm is proposed in the Meta-scheduler Model as shown in Fig. 1. The structure of the cluster consists of one Head Node and 'N' number of computing Nodes as is shown in the Fig. 2. The users will submit their jobs described using Job Submission Description Language (JSDL) specification to the underlying meta-scheduler which in turn falls into the queue of Request Handler component. This queue is called as the request handler queue or ready queue. The structure of meta-scheduler model consist of various component such as Request Handler (RH), Runtime Estimator (RE), Database, Dispatch Manager (DM), Transfer Manager (TM), Execution Manager (EM) and Information manager (IM) as shown in Fig. 3. Here the RH provides the user interface and it maintains the ready queue for obtaining the jobs submitted by different user.

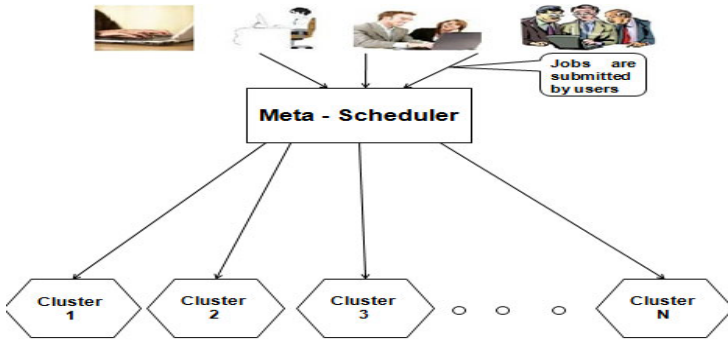


Fig. 1. Meta-Scheduling Model

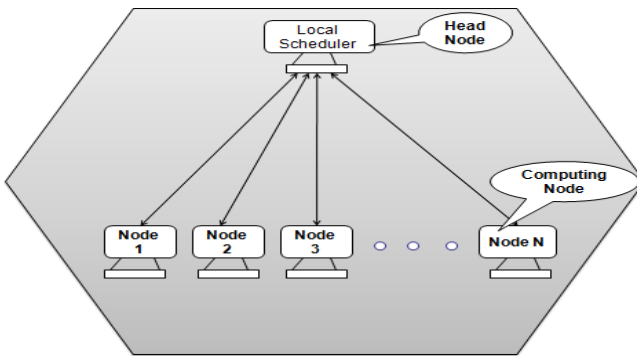


Fig. 2. Structure of the Cluster

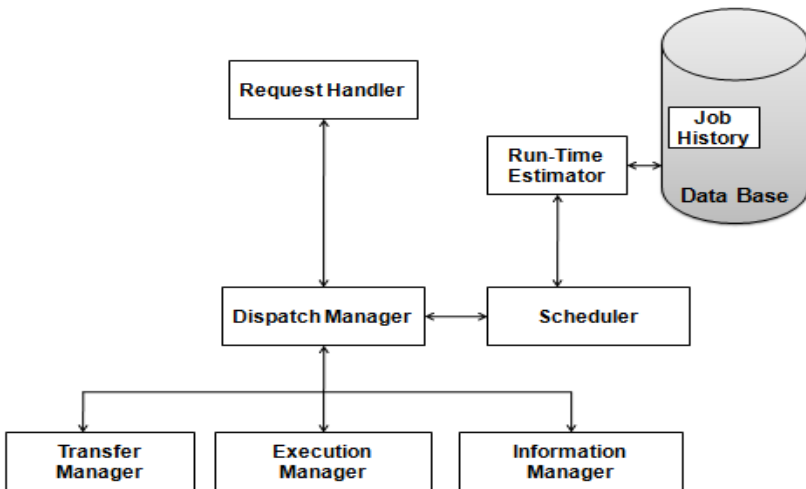


Fig. 3. Structure of Meta-Scheduler

The DM component maintains the time interval for pulling the jobs from the ready queue. For each time interval DM will periodically pulls the job from RH and give it to Scheduler component for appropriate scheduling. Now this scheduler component will exploit the proposed scheduling algorithm to map with the best possible resources available as shown in Algorithm 1. First the algorithm checks for the runtime of the job in the job specification. If it finds the runtime attribute in the job specification, then it will probably finds the appropriate matched resource for job submission and as a result it will migrates the job to the concern resource for execution. In-case if the scheduling algorithm doesn't find any runtime attribute in the job specification, then it will automatically calls the RE component to find the estimated runtime of the job from the Job History. This situation arises since many users submit their jobs with the incomplete job requirement (without specifying the runtime). These types of jobs with incomplete job requirement or with inaccurate runtime will leads to poor scheduling. Most of the cases user may not be aware of the jobs nature, so runtime may not be specified in the job requirement. To handle such a jobs, runtime or execution time is estimated by using the RE component which in turn query the database with some attributes of the jobs to get the runtime information of job from the Job History. The accurate runtime estimation helps to achieve better resource utilization, reduce waiting time, proper resource allocation and satisfying user level Quality of Service.

Finally the scheduler component will finds the matched resource-id and then invoke the DM for dispatching the job to matched resource. The IM will query the Monitoring and Discovery Service (MDS) and sends the resource or host information to the Scheduler. Based on the monitoring interval it keeps track of the host status and updates the host information such as if any new resources are added or removed in the Grid Environment that information are updated periodically. The TM is invoked by the DM with the job-id and the matching resource-id as input. Once it is invoked, the TM creates a remote directory for the given path name as one specified in user specification or jobs input. TM gives the permission rights for the execution of given job in the remote directory. Once this process is over, it informs the DM through messages. The EM is invoked by the DM when the TM completed the creation of directory in the remote host. The DM will dispatch the job for execution and the EM will keeps on monitoring and updating the job status to the scheduler. Finally EM reports the scheduler about the completion of job in case of successful execution and reports the failure in case of job incompleation.

5 Runtime Estimation Aware Scheduling Algorithm

Exploiting of this algorithm in the meta-scheduler will helps to predict the completion time of deadline based jobs in the available resource. The Completion Time of Job 'J_i' in the available 'k' number of resource is computed as shown in equation (1) and (2),

$$CT_{J_i}(R_k) = WT_{J_i}(R_k) + ET_{J_i}(R_k) \quad (1)$$

Where $WT_{J_i}(R_k)$ denotes the Waiting Time of Job 'J_i' in the resource R_k and $ET_{J_i}(R_k)$ represents the Execution Time of the Job 'J_i' in the resource R_k . Here the Waiting Time of the job in particular resource is computed as follows,

$$WT_{J_i}(R_k) = \sum [ET_{WJ_1}(RQ_k), ET_{WJ_2}(RQ_k), \dots, ET_{WJ_m}(RQ_k)] \quad (2)$$

Where $ET_{WJ_1}(RQ_k)$ denotes the Execution Time of Waiting Job ‘ WJ_1 ’ in the Resource Queue RQ_k and the ‘ m ’ indicates the number of jobs waiting in Resource Queue of the resource R_k .

Algorithm 1. Runtime Estimation Aware Scheduling (REAS) Algorithm

```

Begin
Get information about list of jobs waiting in resource queue
for ( all job )
    if ( job requirement has the execution time )
        Compute Completion time of job in all resource
        Identify the resource having minimum Completion time
        Submit the job to selected resource
    else
        Identify the similar types of job from the job history
        Get the execution time of similar jobs present in the Database
        Add the execution time to the job requirement
        Compute Completion time of job in all resource
        Identify the resource having minimum Completion time
        Submit the job to selected resource
End
    
```

6 Experimental Results and Performance Evaluation

In the result phase we have evaluated the result by exploiting five resources with hundreds of dissimilar jobs (deadline and non deadline based). The result of the Proposed Scheduling Algorithm (PSA) in paper [3] is compared with the proposed REAS algorithm and its performance measure is also represented as a graph as shown in Fig. 4. In-order to show exact performance of the proposed algorithm, the graph is

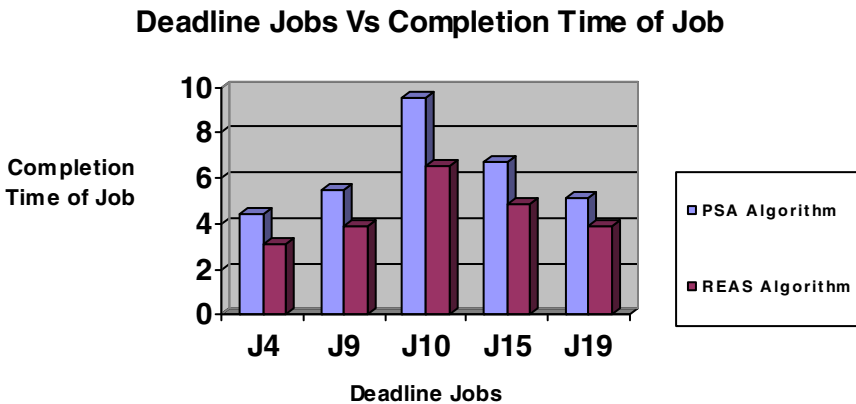


Fig. 4. Performance Evaluation of PSA Algorithm Vs REAS Algorithm

generated for deadline based jobs. In the job submission experimental model there is only five deadline based job such as J4, J9, J10, J15 and J19. From the graph it is clear that the performance of the proposed REAS algorithm works much better than the Proposed Scheduling Algorithm (PSA) in paper [3] with respect to completion time of deadline based jobs.

7 Conclusion and Future Work

In this paper a complete role of meta-scheduler using Runtime Estimation Aware Scheduling algorithm is evaluated with the simulation and traces from real time grid environment as well as using GridSim toolkit. The result obtained with performance evaluation can effectively schedule the job to the underlying resource by giving more priority to deadline based job (interactive job). Hence from the result it indicates that completion time of the priority jobs can be considerably optimized by using Runtime Estimation Aware Scheduling algorithm in the Meta-scheduler. In the future we will work towards the exploitation of Service Level Agreement in the meta-scheduler to provide the Quality of Service in the real Grid Environment.

References

1. Anjum, A., Mc Clatchey, R., Ali, A., Willers, I.: Bulk Scheduling with the DIANA Scheduler. *IEEE Transactions on Nuclear Science* 53(6) (December 2006)
2. Rasooli, A., Mirza-Aghatabar, M., Khorsandi, S.: Introduction of Novel Rule Based Algorithm for Scheduling in Grid Computing Systems. In: *Second Asia International Conference on Modeling and Simulation* (2008)
3. Baghban, H., Rahmani, A.M.: A Heuristic on Job Scheduling in Grid Computing Environment. In: *Seventh International Conference on Grid and Cooperative Computing* (2009)
4. Foster, I., Kesselman, C.: *The Grid: Blueprint for a Future Computing Infrastructure*. Morgan Kaufmann (1999)
5. Guim, I.R.F., Corbalan, J.: Evaluation of Coordinated Grid Scheduling Strategies. In: *11th IEEE International Conference on High Performance Computing and Communications* (2009)
6. Lin, J., Gong, B., Liu, H., Yang, C., Tian, Y.: An Application Demand aware Scheduling Algorithm in Heterogeneous Environment. In: *11th International Conference on Computer Supported Cooperative work in Design* (2007)
7. Li, K.: Job Scheduling for Grid Computing on Metacomputers. In: *19th IEEE International Parallel and Distributed Processing Symposium* (2005)
8. Li, K.: Job Scheduling for Grid Computing on Metacomputers. In: *19th IEEE International Parallel and Distributed Processing Symposium* (2005)
9. Yu, K.-M., Chen, C.-K.: An Evolution-based Dynamic Scheduling Algorithm in Grid Computing Environment. In: *Eighth International Conference on Intelligent Systems Design and Applications* (2008)
10. Ni, L., Zhang, J., Yan, C., Jiang, C.: A Heuristic algorithm for Task Scheduling based on Mean Load. In: *First International Conference on Semantics, Knowledge and Grid* (2006)

11. Cendron, M.M., Westphall, C.B.: A price-based Scheduling for Grid Computing. In: Seventh International Conference on Networking (2008)
12. Rajavel, R., Somasundaram, T.S., Govindarajan, K.: Dynamic Load Balancer Algorithm for the Computational Grid Environment. In: Das, V.V., Vijaykumar, R. (eds.) ICT 2010, Part I. CCIS, vol. 101, pp. 223–227. Springer, Heidelberg (2010)
13. Zhu, Y., Li, M., Weng, C.: Ant algorithm with Execution Quality Based Prediction in Grid Scheduling. In: Fourth China Grid Annual Conference (2009)

Particle Swarm Optimization Algorithm vs. Genetic Algorithm to Solve Multi-Objective Optimization Problem in Gait Planning of Biped Robot

Rega Rajendra and Dilip Kumar Pratihari

Soft Computing Laboratory,
Mechanical Engineering Department
Indian Institute of Technology,
Kharagpur-721 302, India
regaraj@gmail.com, dkpra@mech.iitkgp.ernet.in

Abstract. This paper deals with multi-objective optimization problems in ascending and descending gait planning of biped robot, which has been solved using particle swarm optimization algorithm and genetic algorithm separately. In order to model this problem, two modules of adaptive neuro-fuzzy inference systems have been adopted. Two contrasting objectives, such as power consumption and dynamic balance margin have been considered, and Pareto optimal front of solutions has been obtained.

Keywords: Biped gait, Multi-objective optimization, PSO, GA.

1 Introduction

Current research in robotics aims to build intelligent, energy efficient and dynamically balanced biped robots capable of moving through various terrains, as the situation demands. Intelligence is incorporated in a robot artificially, in the form of adaptive motion (path and/or gait) planner. The robot can be made energy efficient through the optimization of its mechanical structure. Moreover, Dynamic Balance Margin (DBM) of a biped robot can be measured using the concept of Zero Moment Point (ZMP) [1]. Studies are being conducted to develop energy efficient biped robots and their gaits using multi-objective optimization evolutionary algorithms. Lee and Lee [2] generated walking patterns for the best performance of a biped robot using multi-objective evolutionary algorithm. Three contrasting objectives were considered, namely mobility, energy efficiency and stability of a robot to obtain optimal set of solutions for generating walking gaits on flat surface. Niehaus and R'offer [3] used the PSO algorithm for walking gait optimization of a humanoid robot. The biped gait was modeled utilizing a number of parameterizable trajectories to achieve omni-directional walking. The optimized set of walking parameters was successfully implemented on a modified Kondo KHR-1 robot on flat surface. Kim et al. [4] used nonparametric estimation-based PSO for finding the parameters of Central Pattern Generator (CPG). The PSO algorithm was able to efficiently determine CPG parameters for a biped gait.

Most of the studies on biped robots published in the literature are related to their locomotion on flat terrains. However, biped robots work on rough terrains also, such as staircases, and others. In the present work, gait planning problem of a 7-dof biped robot ascending and descending the staircase has been modeled as a multi-objective optimization one. Two modules of adaptive neuro-fuzzy inference system (ANFIS) have been utilized to model gait planning problem of the biped robot. Two conflicting objectives, such as minimization of power consumption and maximization of dynamic balance margin have been considered in the present study, and a GA and a PSO algorithm have been utilized to yield Pareto-optimal front of solutions separately. A comparison on the performances of these two optimization algorithms has also been presented.

2 Mathematical Formulation of the Problem

This study deals with an analysis of a 7-dof (that is, three at hip, two at knee and two at ankles) biped robot ascending and descending some staircases.

2.1 Staircase Ascending

The schematic view of a biped robot ascending staircase is shown in Fig. 1. The mass of each link is assumed to be concentrated at a point on it. For simplicity, the movement of the robot is considered in one direction. During motion, the swing foot of the robot is assumed to follow a cubic polynomial trajectory, as follows: $z = c_0 + c_1x + c_2x^2 + c_3x^3$, where z represents the height of the swing foot from the surface of lower staircase, at a distance x_1 from the starting point; and c_0, c_1, c_2 and c_3 are the coefficients, whose appropriate values are to be determined using the boundary conditions. The hip trajectory is assumed to follow a straight line having a slope equal to that of the staircase. The robot is checked for its dynamic balance using the concept of zero moment point (ZMP). It is said to be dynamically balanced, when the ZMP lies inside the foot support polygon. The position of ZMP with respect to the ankle joint measured in the direction of motion is as follows:

$$x_{ZMP} = \frac{\sum_{i=1}^7 \left(I_i \dot{\omega}_i + m_i x_i \left(\ddot{z}_i - g \right) - m_i \ddot{x}_i z_i \right)}{\sum_{i=1}^7 m_i \left(\ddot{z}_i - g \right)}, \tag{1}$$

where I_i denotes the moment of inertia of i^{th} link ($\text{kg}\cdot\text{m}^2$), $\dot{\omega}_i$ is the angular acceleration of link i in (rad/s^2), m_i denotes the mass of i^{th} link (kg), (x_i, z_i) is the coordinate of i^{th} lumped mass, g is the acceleration due to gravity (m/s^2), \ddot{z}_i is the acceleration of link i^{th} in z-direction (m/s^2), \ddot{x}_i is the acceleration of link i^{th} in x-direction (m/s^2).

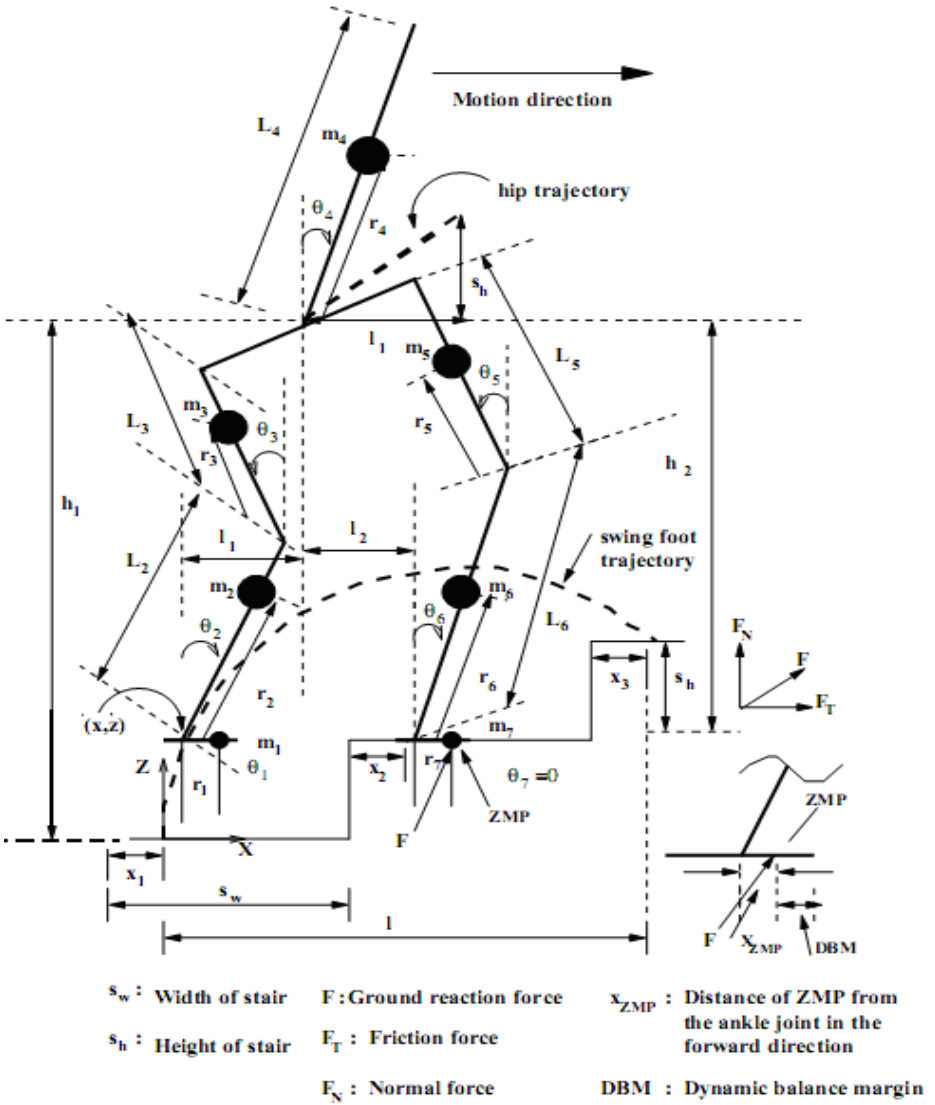


Fig. 1. A schematic view of a biped robot ascending the staircase

Dynamic balance margin (DBM) is calculated as the distance of ZMP from the boundary of support polygon as follows:

$$DBM = \left(\frac{L_7}{2} - |x_{zmp}| \right), \quad (2)$$

where L_7 is the length of the supporting foot and x_{ZMP} represents the distance of ZMP from the ankle joint in the direction of motion. The D-H parameters is used

to generalized parameters (q_i) and joint angles (θ_i) are as follows: $q_1 = \theta_1$; $q_2 = (90 - \theta_2)$; $q_3 = (\theta_2 - \theta_3)$; $q_4 = (\theta_3 - \theta_4)$; $q_5 = (\theta_4 - \theta_5)$, $q_6 = (\theta_5 - \theta_6)$; $q_7 = (\theta_7 - (90 - \theta_6))$. The generalized angles are assumed to follow fifth-order polynomials in order to ensure their smooth variations, as follows: $q_i(t) = a_{i0} + a_{i1}t + a_{i2}t^2 + a_{i3}t^3 + a_{i4}t^4 + a_{i5}t^5$, where $i=1,2\dots n$ joints and $a_{i0}, a_{i1}, a_{i2}, a_{i3}, a_{i4}, a_{i5}$ are the coefficients, whose values are to be determined using some known conditions. The angular velocity and acceleration can be determined by differentiating $q_{i(t)}$ with respect to time once and twice, respectively. Torque (τ) required at each joint of the robot for its locomotion has been determined using Lagrange formulation [5]. The amount of power consumed by i^{th} joint can be calculated as the product of motor torque and angular velocity. If the amount of heat loss of the motor is considered [6], the average power consumption over a cycle of time period T , is calculated as follows:

$$P_i = \frac{1}{T} \sum_{i=1}^n \int_0^T \left(\left| \tau_i \dot{q}_i \right| + K \tau_i^2 \right) dt \quad (3)$$

where K is a constant, whose value has been assumed to be equal to 0.025.

The constrained multi-objective optimization problem has been formulated as follows:

Minimize average power consumption

$$P_i = \frac{1}{T} \sum_{i=1}^n \int_0^T \left(\left| \tau_i \dot{q}_i \right| + K \tau_i^2 \right) dt,$$

and Minimize 1/DBM

$$1/DBM = 1 / \left(\frac{L_7}{2} - |x_{zmp}| \right)$$

Subject to

$$\Delta \tau_{ij} = \Delta \tau_{specified}; \text{ and } r_1^{\min} < r_1 < r_1^{\max}, r_2^{\min} < r_2 < r_2^{\max}, r_3^{\min} < r_3 < r_3^{\max}, r_4^{\min} < r_4 < r_4^{\max}, m_4^{\min} < m_4 < m_4^{\max}.$$

The parameters: r_1, r_2, r_3 and r_4 denote the mass center positions of first, second, third and fourth links, respectively and m_4 represents the trunk mass. T indicates cycle time and n represents the number of joints. Due to symmetry of the biped robot, r_5, r_6 and r_7 have been kept equal to r_1, r_2 and r_3 , respectively. Here, $\Delta \tau_{ij}$ represents the change in torque of i^{th} joint at j^{th} time interval. It has to be less than some pre-specified value to ensure smooth motion. The gait planning problem of the robot for descending the staircase has also been solved using the similar procedure.

3 Proposed Algorithms

Gait planning problems has been modeled using two ANFIS [7, 8] modules for the biped robot ascending and descending the staircase, as shown in Fig. 2. The ANFIS model involves two inputs and outputs each. Each variable has been modeled using four linguistic terms: L (low), M (medium), H (high), VH (very high). The inputs for two variables of the first module vary in the range of 0.03 to 0.105 (m); and for the second module, the first and second variables vary in the ranges of $(-10^0$ to $11^0)$ and $(-10^0$ to $15^0)$, respectively. Each ANFIS module generates output rules in the form of $Y_{ji} = a_i x_1 + b_i x_2 + c_i$, where $j=1,2$ and $i=1,2,\dots,16$ for each output. The range of variation for a, b, c is kept equal to 0.0 to 1.0. Fig. 3 displays the schematic view of an ANFIS involving two inputs and outputs each.

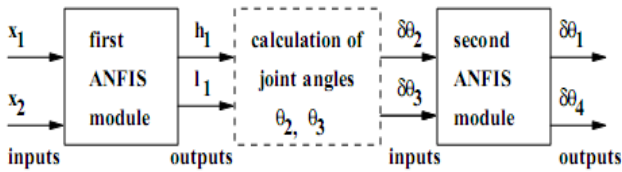


Fig. 2. A schematic view of two ANFIS modules considered for biped gait planning

3.1 Genetic Algorithm

Genetic Algorithm (GA) is a population-based probabilistic search technique based on the principle of natural genetics. The evolution occurs through reproduction, crossover, and mutation. Randomly generated population of solutions are sorted for non-domination (rank or fitness), and the crowding distance is calculated between solutions. Non-dominated solutions are stored for comparison for next generation and dominated solutions are deleted. GA operators are used to create next population of solutions. The GA has been used for solving multi-objective optimization problem also. Non-dominated Sorting Genetic Algorithm (NSGA-II) [9] is one of such examples, which is capable of handling constrained optimization problems.

3.2 Particle Swarm Algorithm

Particle swarm optimization (PSO) [10] is based on swarm behavior in searching food (objective) in a (d - dimensional) search space. Particle is defined by position vector (x_{id}) movement or velocity vector (v_{id}). A multi-objective PSO (that is, MOPSO-CD) [11, 12] incorporates the mechanism of crowding distance of NSGA-II into the PSO. The concept of crowding distance together with mutation operator maintains the diversity of non-dominated solutions. The Pbest (P_{id}) and Gbest (P_{gd}) are sorted for non-dominated front of solutions through the comparison of fitness values of the particles.

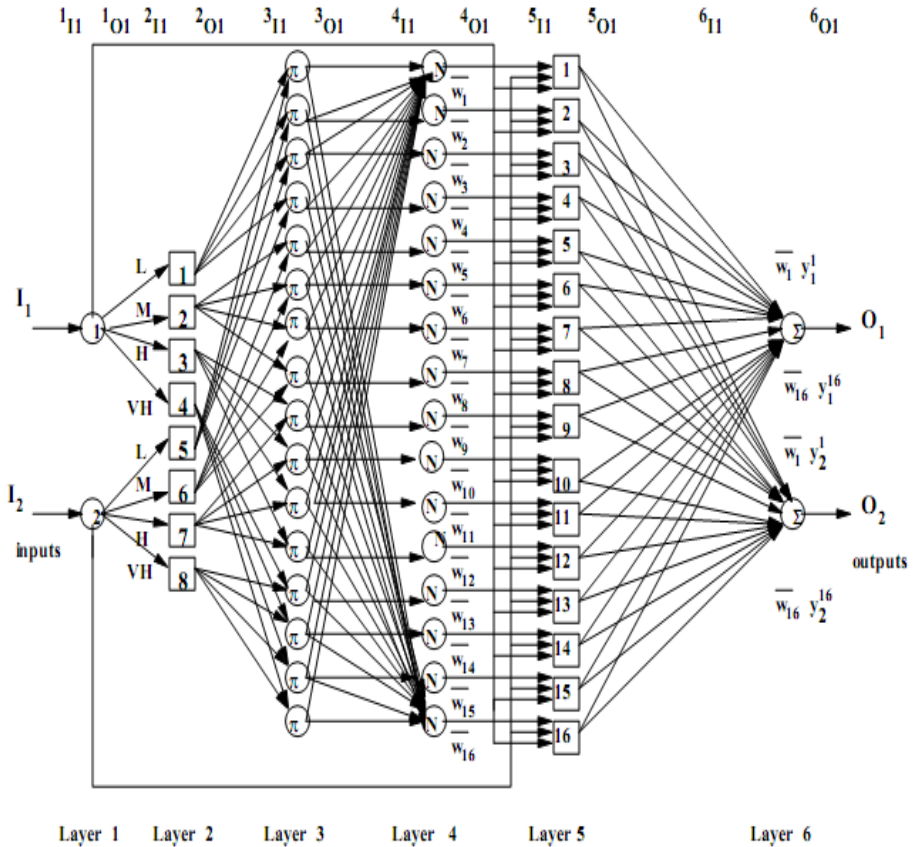


Fig. 3. A schematic view of an ANFIS for two inputs and two outputs used in gait planning

The new velocities and positions of the particles are calculated as follows:

$$v_{id}(t+1) = W \times v_{id}(t) + c_1 \times rand(.) \times (P_{id} - x_{id}(t)) + c_2 \times Rand(.) \times (P_{gd} - x_{id}(t)), \quad (4)$$

$$x_{id}(t+1) = x_{id}(t) + v_{id}(t+1), \quad (5)$$

The coefficients of acceleration are c_1 and c_2 and W is inertia weight at t^{th} time instant. The fitness values of PSO and GA solutions have been determined corresponding to the said two objectives, A penalty function approach has been adopted in order to penalize a solution, if there is a violation of functional constraint. In the present problem, a total of 201 variables have been considered. The algorithms will try to determine the solution corresponding to the maximum dynamic balance margin of the robot but at the expense of minimum power.

4 Results and Discussion

The biped robot consists of links having the masses (in kg): $m_1 = m_7 = 0.5; m_2 = m_6 = 2.0; m_3 = m_5 = 5.0$ and m_4 is varied in the range of 10.0 to 50.0 kg. The links are assumed to have the lengths (in m) as follows: $l_1 = l_7 = 0.06, l_2 = l_6 = 0.34, l_3 = l_5 = 0.30, l_4 = 0.6$. The cycle time t has been assumed to be equal to 5.0 seconds. The maximum velocity of the swing foot has been considered to be equal to 0.056 m/s. In simulations, the values of, $r_1^{\min}, r_1^{\max}, r_2^{\min}, r_2^{\max}, r_3^{\min}, r_3^{\max}, r_4^{\min}, r_4^{\max}, r_5^{\min}, r_5^{\max}, r_6^{\min}, r_6^{\max}, r_7^{\min}, r_7^{\max}$ and m_4^{\min}, m_4^{\max} have been set as 0.01, 0.02, 0.1, 0.32, 0.1, 0.28, 0.1, 0.54, 0.01, 0.02, 0.1, 0.32, 0.1, 0.28, 10.0 and 50.0, respectively. Computer simulations are carried out on a P-IV PC.

4.1 Results of Constrained Optimization

Constrained optimization using penalty function has been adopted to solve problems related to ascending and descending the staircases using the GA and PSO algorithm, separately. The following GA-parameters have given the best results: crossover probability $p_c = 0.8$, mutation probability $p_m = 0.00505$, maximum number of generations = 100 and population size = 100 for the ascending and descending cases. Similarly, the following PSO-parameters obtained through a careful study are seen to yield the best results: number of runs=100, swarm size =100 for both the ascending and descending cases. Fig. 4 displays the Pareto-optimal fronts of solutions obtained by the GA and PSO algorithm for the ascending and descending cases. In both cases, PSO algorithm has outperformed the GA. It has happened due to the reason that PSO algorithm can carry out both the global and local searches simultaneously, whereas the GA is poor in its local search.

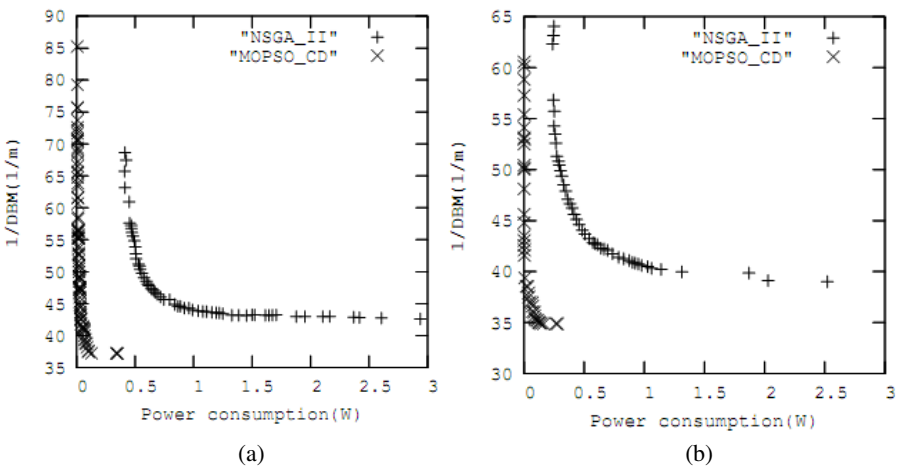


Fig. 4. Pareto-optimal fronts of solutions for (a) ascending, (b) descending problems

5 Concluding Remarks

Gait planning problems of a 7-dof biped robot ascending and descending some stair-cases are considered with two conflicting objectives, namely minimum power consumption and maximum dynamic stability margin. The constrained optimization problems have been solved using the GA and PSO algorithm separately to obtain Pareto-optimal fronts of solutions. The findings of the study are in tune with the general experience of human beings. The PSO has performed better than the GA, as the former carries out both the global and local searches simultaneously, whereas the latter is a potential tool for global search only. The obtained Pareto-optimal fronts of solutions may help the designer to select some appropriate optimal solutions depending on the requirements.

References

1. Vukobratovic, M., Frank, A.A., Juricic, D.: On the stability of biped locomotion. *IEEE Trans. on Biomedical Engineering* 17(1), 25–36 (1970)
2. Lee, J.Y., Lee, J.J.: Optimal walking trajectory generation for a biped robot using multi-objective evolutionary algorithm. In: *Proc. of IEEE Control Conference, Melbourne, Australia*, vol. 1, pp. 357–364 (2004)
3. Niehaus, C., Röfer, T., Laue, T.: Gait optimization on a humanoid robot using particle swarm optimization. In: *Proc. of the Second Workshop on Humanoid Soccer Robots, IEEE-RAS, Intl. Conf. on Humanoid Robots, Pittsburgh, PA, USA* (2007)
4. Kim, J.J., Lee, J.W., Lee, J.J.: Central pattern generator parameter search for a biped walking robot using nonparametric estimation based particle swarm optimization. *Intl. J. of Control, Automation and Systems* 7(3), 447–457 (2009)
5. Fu, K.S., Gonzalez, R.C., Lee, C.S.G.: *Robotics: Control, Sensing, Vision, and Intelligence*. McGraw-Hill Inc. (1987)
6. Nishii, J., Ogawa, K., Suzuki, R.: The optimal gait pattern in hexapods based on energetic efficiency. In: *Proc. of the 3rd Intl. Symp. on Artificial Life and Robotics, Hong Kong, October 29 - November 01*, pp. 106–109 (1998)
7. Jang, J.R.: ANFIS: Adaptive-network-based fuzzy inference system. *IEEE Trans. on Systems, Man and Cybernetics, Part B* 23(3), 665–685 (1993)
8. Pratihar, D.K.: *Soft Computing*. Narosa Publishing House, New Delhi (2008)
9. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. on Evolutionary Computation* 6(2), 182–197 (2002)
10. Clerc, M., Kennedy, J.: The particle swarm-explosion, stability and convergence in a multi-dimensional complex space. *IEEE Trans. on Evolutionary Computation* 6, 58–78 (2002)
11. Raquel, C.R., Naval Jr., P.C.: An effective use of crowding distance in multi-objective particle swarm optimization. In: *Proc. of Genetic and Evolutionary Computing Conference (GECCO 2005), Washington DC, USA, June 25-29*, pp. 257–264 (2005)
12. <http://www.particleswarm.info> (accessed from January to July 2010)

Preventing Forest Animals from Train Accidents Using Outlier-Analysis Algorithm in WSN

V.P. Jayachitra and Sumalatha Ramachandran

Madras Institute of Technology,
Anna University, Tamilnadu, India
jayachitravp@annauniv.edu,
sumalatha.ramachandran@gmail.com

Abstract. In the recent decade the incidence of animal fatalities involving trains has remained high in the country. According to recent survey by Wildlife Trust of India (WTI), 72 animals are dying each year due to collision with speeding trains. Its high time we protect the lives of endangered species of animals. Though railway authorities ordered the drivers to reduce the speed of the trains inside forest areas, it does not have any fruitful results so far. We need a mechanism to alert the animals from crossing railway tracks when the train is approaching near. This paper proposes a simple and efficient technique which alerts animals about speeding trains. Unlike other techniques, our proposed mechanism does not need human intervention for operation.

Keywords: Data mining, Forest animals, train accidents, sensors, vibration.

1 Introduction

In the past decade, the tracks running through protected forest areas were converted from meter gauge to broad gauge, allowing the trains to run faster than before. Further, movement of goods train has also increased in the last four years. Most of the animals in the forest try to cross the railway track and they eventually die after clashing with the speeding train. This scenario is prevalent in Eastern and North-eastern states where railway tracks are running through forest areas to transport raw materials like coal and iron ore from the mines to the industry. Goods trains pass through the prime forest area at odd hours, including the period after evening when animals go out searching for food. Elephants are the major victims in train accidents. The death of scores of elephants on a train track in North Bengal has hit the news recently. To preserve the ecology of the earth and to protect the animals from being extinct, we need a mechanism to alert animals from crossing the tracks when trains are approaching nearby. This can be done by the deployment of sensors in railway tracks. Sensor networks are highly distributed networks of small, lightweight wireless nodes deployed in large numbers to monitor the environment. Sensors usually rely only on their battery for power, which in many cannot be recharged and also requires human intervention for replacing. Hence, the energy of the nodes is a major criterion that decides the network longevity. This makes energy consumption a critical factor in

the design of WSN. So energy consumption should be minimized for increasing the lifetime of the sensor nodes. We focus on how sensor nodes can be deployed to alert animals in such a way that the mechanism involves less power consumption.

2 Survey of Animal Deaths in Train Accidents

One of the major victims in the train accidents is the Asian elephant. Elephant deaths in railway accidents have been reported from all elephant range states in India, with more than 170 train-hit deaths recorded since 1987. Nearly 90% of these deaths in the past 2 decades were recorded in Assam, West Bengal, Uttarakhand and Jharkhand. The toll of elephant deaths due to train accidents in the year 2008 to 2010 is 36. More than 160 elephants have been killed in train accidents in India since 1987. The state of West Bengal accounts for about 26% of the total, second only to Assam which accounts for 36%. In November 2010, at least seven elephants were mowed down by a passenger train in Upper Assam before the engine derailed. An inter-city passenger train coming from Ledo to Dibrugarh hit a small herd of elephants sitting on the tracks at around 5.15 p.m. leading to the deaths of at least two adult elephants and four calves. In an earlier incident, an elephant calf was fatally injured in a train accident at Mahananda wildlife sanctuary in West Bengal. The elephant was hit by a goods train while crossing a railway line on the outskirts of the sanctuary in Siliguri, in search of food in the nearby fields. Forest officials said there are three main migratory herds of elephants, which walk the forests of north Bengal while migrating to and from Assam and Nepal. The railway tracks pass through migratory corridors at 15 places. Since 2000, when locomotives began using the 120-km Siliguri-Alipurduar rail route, there have been 10 elephant deaths. The forest department has no record of injuries to animals in rail accidents. Elephant deaths due to train hits are not only restricted to the North-East. A pair of elephants was allegedly hit by the Pune-Howrah Azad Hind Express between Posaita and Manoharpur stations about 120 km. from Jamshedpur, Jharkhand in November, 2001. The elephants subsequently died. In May 2001, a young cow elephant, around 16 years old was hit by the Mussoorie Express. The accident happened on the rail track running through the Motcihur range of the Rajaji National Park. This was the latest in a series of accidents on this track, which has claimed the lives of elephants, leopards and deer.

2.1 Previous Techniques

As a solution to prevent animals from train accidents, various monitoring systems, security patrolling, whistle blowing techniques are used so far. All these techniques need human power or human intervention for their operation. But our proposed technique with sensors requires human intervention only during the deployment of sensors in railway. We also focus on reducing energy consumption in the sensor nodes so that lifetime of the network is increased. For efficient communication of sensor nodes we employ clustering strategy. Clustering architecture with cluster heads assuming key roles of routing is the widely followed model these days. More recently coverage time optimization is implemented via an

analytical model [1]. This coverage time optimization involves routing aware optimal cluster planning and clustering aware optimal random relay. The case of both deterministic deployment and random deployment is considered. In deterministic deployment clustering is based on swapping and in random deployment clustering is done in a ring like fashion. But there are assumptions like the sensing field is assumed to have a uniform area and be symmetric, which is not practically possible. Swapping can reduce energy consumption inside the cluster but no such effect on inter-clustering is achieved. To overcome the failure of a single cluster head, a method in which each node in the cluster acts as CH, on a rotation basis is proposed [1]. According to this method, the node with the biggest residual energy is selected as the cluster head. Since the cluster head keeps on changing it requires additional overhead in propagating the election of a node as head to other nodes which consumes more energy each round. The placement of CH's also plays a role in reducing energy consumption. The optimal placement of cluster heads is proposed [6] in which two nodes communicate only when the distance between the two nodes falls within the communication range. Energy consumption is reduced but free transmission is restricted.

In our work, we propose significant reduction in energy by means of more efficient clustering algorithm based on outlier analysis. In general outlier analysis is used to detect malicious sensor nodes [10]. This is done by recording the normal behavior of sensor nodes and comparing them with any new behavior to see if the node is malicious. But in our approach we use outlier analysis to elect cluster head.

3 Implementation Scenario

With increase in urbanization, forest lands are shrinking in size. Thus many rare species of animals are becoming extinct. It's high time we protect the lives of endangered species of animals. We need a mechanism to alert the animals from crossing railway tracks when the train is approaching near.

Deploying sensors to alert the animals by physical means can be effective. But this application requires of longevity of the sensor nodes. Hence our algorithm can be better suited for this application.

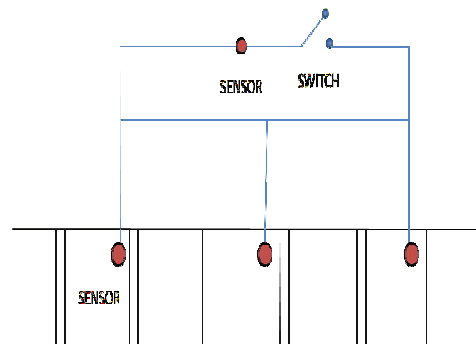


Fig. 1. Deployment of sensor nodes in railway tracks

The sensor nodes are placed in railway tracks as shown in figure, at the rate of 10 sensors per km. The sensors in a km are attached to an electric circuit. The circuit consists of a sensor and a switch. When the wheels of the train move over the tracks, vibrations are produced. The vibrations travel at a velocity higher than that of the velocity of the moving train. When the train is at a distance of 1 km from a position say x , the vibration at points near x will be in the range of 1300-1500 Hz [18]. The places far away from x will measure vibration in the range of 40 Hz. The sensors in the track measure the vibration produced. When the measured vibration reaches the threshold of 1300 Hz the sensor sends a signal to the sensor in the nearby circuit. On receiving this signal, the sensor in the circuit pushes the switch to make a connection. Current in the range of 1 volt flows in the closed circuit. The rail lines which are made of iron, conduct current. When the animals try to cross the track, they will feel a mild shock. Thus they step back and move without crossing the track further. The electric power to the switch can be obtained from nearby electric post or even a battery is sufficient.

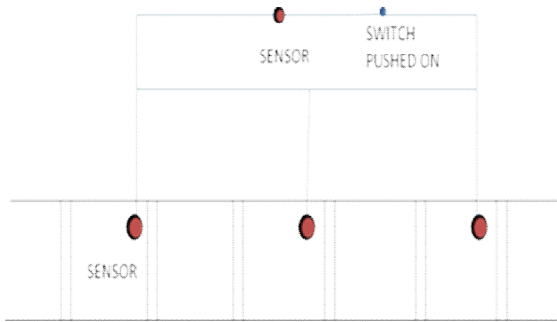


Fig. 2. Scenario When Train is Within 1 km

This method is easy to implement and prevents the animals from dashing against speeding trains without causing any harm to them. This process requires the sensor nodes to retain their life long. Hence our algorithm can be implemented to communicate to the sensor in the circuit (which is the sink here).

4 Network Scenario

The network scenario is proposed in four phases. Phase 1 deal with layering where information about position of all nodes is obtained. Phase 2 deals with electing leader nodes and forming clusters with leader nodes as cluster heads and phase 3 deals with electing cluster heads based on outlier analysis. Phase 4 explains how about how aggregation is implemented at the cluster heads and data is sent to the base station.

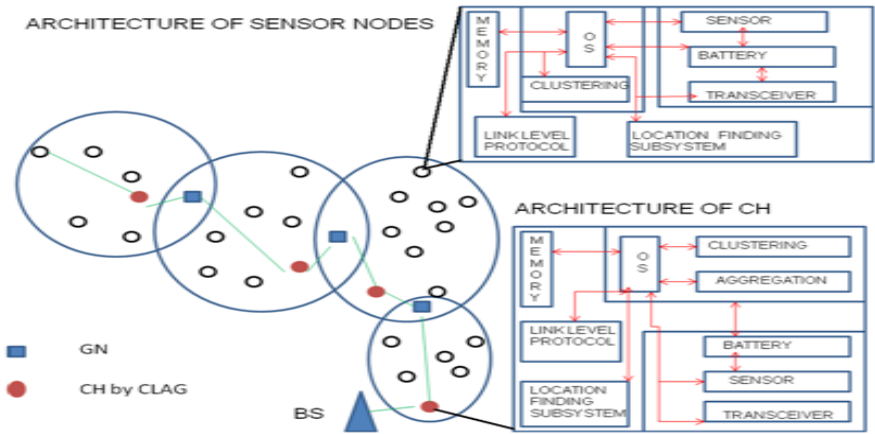


Fig. 3. Network Scenario

A. Phase 1- Layering

Initially the base station does not know the position of the sensor nodes. The base station needs a mechanism to obtain the geographical position of the sensor nodes based on which the cluster heads are chosen. Because of the limited power and range of the sensor nodes that are located far away from the base station, they cannot directly send their location information to the base station, doing so will completely drain their battery power, which makes the entire network to fail. Thus we go for a layering approach. The nodes which can directly transmit data to the base station form the first layer. Base station sends a query message with its ID to all sensor nodes in the first layer. On receiving this message, the sensor nodes send a message containing their ID, position information and the set of all neighboring nodes (which are in its range) to the base station. Upon receiving this message from all layer 1 nodes, the base station chooses a node as Layer head (LH) for layer 1 and unicasts message containing LH information to that elected node. A node which is at mean distance to the base station and which has majority of next layer sensor nodes as its neighbor is chosen as LH. Now the LH of layer1 broadcasts query message to the layer2 nodes. On receiving this message, the sensor nodes send a message containing their ID, position information and the set of all neighboring nodes (which are in its range) to the LH of layer1, which in turn forwards the message to the base station. LH of layer1 elects the LH of layer2. Now LH of layer2 forwards query message to all layer3 nodes collects the reply message from them and forwards it to the LH of layer1. This layering approach is followed until the entire topology of nodes is covered. The leaders in each layer are responsible for forwarding the query message from base station to the nodes in the next layer and collect data from the nodes in the second layer and forward the data to the previous LH. Whenever a node receives a query message with the base station id, the node uses its location finding subsystem to find the position of the node as a function of (x, y) co-ordinates. Later it sends this (x, y) co-ordinate along with its id and neighboring node information to the leader node from which it received the forwarded request.

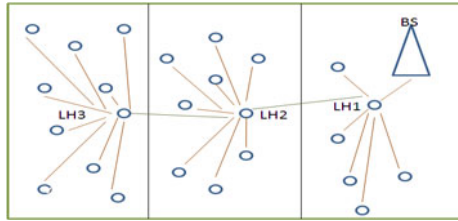


Fig. 4. Phase 1- Layering

B. Phase 2- Cluster Formation

The base station contains all the information about the sensor nodes such as the ID, position and the neighboring nodes in range for each node. Based on this information it chooses a node which has maximum coverage as leader node. The base station finds dominant set such that the chosen set of sensor nodes cover all the sensor nodes in that area. Thus in each layer a node is selected as leader node. The leader node later helps in electing cluster heads based on outlier analysis in the next phase. The base station unicasts the message to the leader nodes informing they are elected. Now the leader nodes start forming a cluster. At present the leader nodes act as cluster heads and initiate cluster formation process. The leader nodes send a message with their identity stating that they are chosen as leader nodes by the base station to all the sensor nodes which are in its vicinity. On receiving this message, other sensor nodes send their position with a join confirmation message to leader nodes which are within the range of the node. A cluster with the leader node as cluster head is formed. The leader nodes maintain local topology information about all sensor nodes in the cluster. If a node receives message from more than one leader node, then this node lies in close vicinity to the two clusters. Such nodes are called gateway nodes. These gateway nodes are the key nodes for inter-cluster communication.

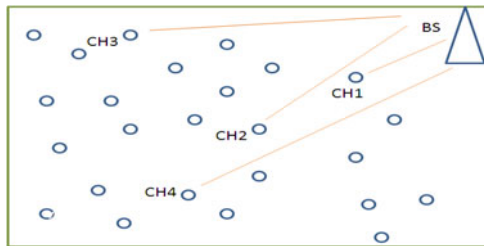


Fig. 5. Phase 2-Cluster formation

C. Phase 3-Cluster Head Selection

The leader node maintains local topology. When it receives position information from each sensor node in the cluster, it calculates the distance of all the nodes. Based on this information, the outlier node (distance based outlier analysis) is then chosen as cluster head. Now leader node sends elected message to the cluster head. On receiving this message cluster head broadcasts its identity to all other nodes in the cluster. Now all further communication takes place through the cluster heads.

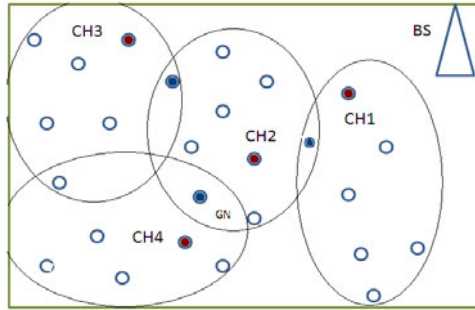


Fig. 6. Phase 3- Cluster Head selection

D. Phase 4- Routing and Aggregation

In this phase intra-cluster routing, aggregation and inter-cluster routing is established. The base station broadcasts the initial timer value to all cluster heads. This timer value is based on Semi-Markov decision process model. Using this model, the timer value is derived based on a generated random sequence of network traffic. When sensor nodes detect an event or measure a value, they send the data to the cluster head. The cluster heads receive data from the sensor nodes and are involved in aggregating the data. The maximum limit for aggregation can be set based on the maximum message size the sensor node can transit. The cluster head carries on aggregation only within the timer value. Throughout this time the cluster head receives data from the sensor nodes within range and aggregates the data for transmission. The CH keeps track of maximum message size each time it aggregates the data. It also saves the frequency information that is the frequency with which the sensor nodes send data. This information is used for aggregation analysis. If the timer is expired or maximum message size is attained, the cluster head goes to transmit node and now transmits aggregated data to the gateway nodes. While transmitting aggregated data, if CH receives any more messages it starts buffering them until it finishes current transmission. After transmitting, CH goes into aggregation mode. While in aggregation mode, the CH aggregates data and also begins analyzing the frequency information saved before by applying pattern mining algorithm. The algorithm detects the frequency at which sensor nodes transmit data and fixes a time based on network traffic as aggregate timer. If the aggregate timer value computed is not equal to the previous aggregate timer value, the cluster head resets the timer value. The gateway nodes, when received data from the cluster head transmits data to the nearest sensor node in the adjacent cluster. Now the sensors in that cluster take care of routing which is similar to intra-cluster routing. Routing proceeds in this fashion and ultimately data is handed over to the base station. Thus data is transmitted to the base station.

Algorithm for Routing and Aggregation:

```

/* At sensor nodes */
  For each node i in network
    If data sensed
      datarate = RAND(50)
      data = [ data sensed + datarate ]
      id_src = id of node i
      id_dest = ch_id
      SEND(data) to id_dest
/* At cluster heads */
Procedure ROUTE (msg data, cluster_head i)
{
  For each cluster_head j in network
    max = 0
    distance = sqrt{ [xpos(j) -xpos(i)]2 +[ypos(j) - ypos(i)]2
  }
    If max > distance
      max = distance
      nexthop = j
    End If
  End For
  id_dest = id of nexthop
  SEND(data) to id_dest
}
For each cluster head i in network
{
  Counter = 5 ms
  Set Timer
  ag_data = " "
  If msg_rcvd
    retrieve datarate from msg
    If datarate < channel_bandwidth and Timer < Counter
      ag_data = ag_data + data_rcvd
    Else If Timer > Counter
      ROUTE (ag_data, i)
      Reset Timer
    Else if datarate > channel_bandwidth
      ROUTE (data_rcvd, i)
}

```

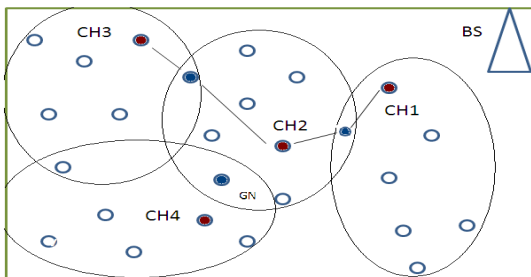


Fig. 7. Phase 4-Routing and aggregation

5 Results

We have simulated our architecture in OMNET simulator. The simulation scenario consisted of 30 sensor nodes and a base station. Let(x, y) denote the position of a sensor. We know that Energy consumption by a node is directly proportional to the distance of transmission and the size of the message transmitted.

$$E_{consump} \propto Distance \quad \dots >$$

$$I E_{consump} \propto size \quad \dots > II$$

But the size of data transmitted by a sensor network is very less, ie sensors will transmit either a bit for signaling occurrence of an event or it will transmit a few bits. Thus we do not consider II in our further proceedings. Considering I, Reducing distance reduces the energy consumed by the node. Thus we focus on proving that outlier analysis based clustering reduces energy consumption of the node. Let us fix that a sensor node needs energy 1 joule

* distance to send message to node which is at its direct transmission range. The transmission range of a 2400Hz sensor is 3 meters. For simulation scenario, we scale it to 3 cm, if we follow the traditional k-means clustering algorithm, then location of the sensors is at the mean distance from the other sensors in the cluster.

$$distance = \sqrt{\{ [xpos(j) - xpos(i)]^2 + [ypos(j) - ypos(i)]^2 \}}$$

Thus in our simulation scenario base station is at (35, 50). The position of sensor nodes given here are scaled down from our simulation scenario. In real time implementation, the distance can be measured in metres.

Table 1. Positions Of Cluster Heads And Gateway Nodes

Position of gateway nodes	CH position by k-means algorithm	CH position by outlier analysis
(42,220)	(45,250)	(45,240)
(46,180)	(40,200)	(45,200)
(50,155)	(48,160)	(53,155)
(60,112)	(55,110)	(60,112)
	(70,80)	(80,85)

In the next step, the energy required by sensor nodes for selecting cluster heads is calculated. We choose five nodes and calculate the energy spent by them.

Table 2. Energy Spent To Elect CH

Sensor node position	Energy consumed to elect CH (in joules)	
	By k-means	By outlier analysis
(23,140)	3.4	3.49
(32,12)	4.1	4.53
(44,150)	3.6	3.62
(12,230)	5.21	5.34
(20,90)	2.11	2.31

The table shows that energy spent by the sensor nodes to elect cluster heads based on outlier analysis is marginally higher than that required by k-means algorithm. But this depicts the intra-cluster scenario, where the sensor nodes may not be active all the time. We mainly focus on reducing energy consumption of cluster heads which are active throughout. Now, energy required by cluster heads to communicate to sink is considered.

Table 3. Energy Spent To Communicate To Sink

Cluster heads in the network	Energy consumed to send message to sink (in joules)	
	By k-means	By OAPM
CH1	30.1	20.2
CH2	20.8	14.2
CH3	5.3	3.5
CH4	5.4	3.6
CH5	49.09	30.92

From our simulation, we arrive the result that our algorithm consumes energy 0.67 times lesser than k-means clustering.

The following graphs depict the residual energy available at the sensor nodes after 25 message transmissions to the cluster head.

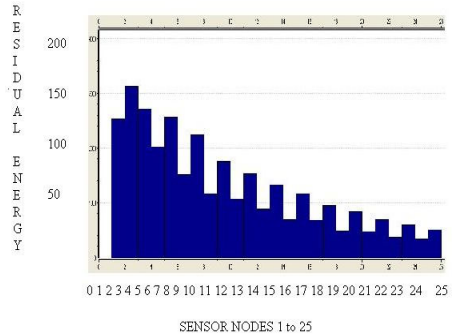
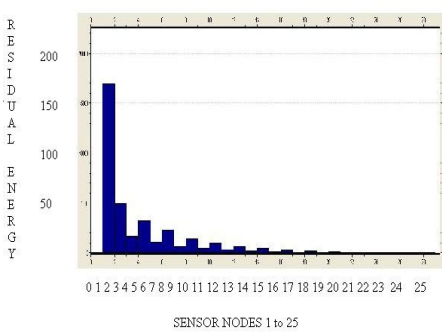


Fig. 8. Residual Energy of Sensor Nodes in K-Means

Fig. 9. Residual Energy of Sensor Nodes in OAPM

6 Conclusion

We have proposed an efficient solution for preventing forest animals from train accidents, which is the present day need. We have prescribed a clustering algorithm which implements the outlier analysis technique which reduces the inter-cluster distance and hence conserves energy at the cluster heads. The results show that significant amount of energy is conserved in the cluster heads by using the clustering algorithm.

References

1. Shu, T., Krunz, M.: Coverage-Time Optimization for Clustered wireless Sensor Networks: A Power-Balancing Approach. *IEEE Transactions on Networking* 18(1) (February 2010)
2. Ye, Z., Abouzeid, A.A., Ai, J.: Optimal Stochastic Policies for Distributed Data Aggregation in Wireless Sensor Networks. *IEEE Transactions On Networking* 17(5) (October 2009)

3. Xiang, M., Luo, Z., Wang, P.: Energy efficient intra-cluster data gathering in wireless sensor networks. *Journal Of Networks* (March 2010)
4. Ren, M., Guo, L.: Mining recent approximate frequent items in wireless sensor network. In: *International Conference on Fuzzy Systems and Knowledge Discovery* (2009)
5. Sin, I., Lee, J.: Performance analysis according to the change of cluster size in large scale wireless sensor networks. *International Journal of Computer Science and Network Security* (April 2009)
6. Pandey, S., Agarwal, P., Dong, S.: *On Performance of Node Placement Approaches for Hierarchical Heterogeneous Sensor Networks*. Springer, Heidelberg (2008)
7. Boukerche, A., Martiryosan, A., Pazzi, R.: *An Inter-cluster Communication based Energy Aware and Fault Tolerant Protocol for Wireless Sensor Networks*. Springer, Heidelberg
8. Wafra, M., Daher, W., Al Azar, H.: A Sensor Network Data aggregation technique. *International Journal of Computer Theory and Engineering* (1) (April 2010)
9. Bonivento, A., Fischione, C., Neechhi, L.: System level design of wireless sensor networks. *Journal of Latex class files* (6) (January 2007)
10. Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: A Survey on sensor networks. *IEEE Communication Magazine* 40(8), 102–114 (2002)
11. Amis, A.D., Prakash, R.: Load-balancing clusters in wireless ad hoc networks. In: *Proc. 3rd IEEE Symp. Application-Specific Syst. Software Eng. Technol.*, Richardson, TX, pp. 25–32 (2000)
12. Amis, A.D., Prakash, R., Vuong, T.H.P., Huynh, D.T.: Max-mind-cluster formation in wireless ad hoc networks. In: *Proc. IEEE INFOCOM*, Tel Aviv, Israel, pp. 32–41 (2000)
13. Baker, D.J., Ephremides, A.: The architectural organization of a mobile radio network via a distributed algorithm. *IEEE Transactions on Commuication* COM-29(11), 1694–1701 (1981)
14. Bandyopadhyay, S., Coyle, E.J.: An energy efficient hierarchical clustering algorithm for wireless sensor networks. In: *Proc. IEEE INFOCOM*, San Francisco, CA, vol. 3, pp. 1713–1723 (2003)
15. Bandyopadhyay, S., Coyle, E.J.: Minimizing communication costs in hierarchically-clustered networks of wireless sensors. *J. Computer Networks* 44(1), 1–16 (2004)
16. Boyd, S., Vandenberghe, L.: *Convex Optimization*. Cambridge Univ. Press, Cambridge (2004)
17. Chiasserini, C.F., Chlamtac, I., Monti, P., Nucci, A.: An energy efficient method for nodes assignment in cluster-based ad hoc networks. *J. Wireless Netw.* 10(3), 223–231 (2004)
18. Esveld, C., De Man, A.: *Use of Railway Track Behaviour for Design and Maintenance*
19. Singh, A.K., Kumar, A., Mookerjee, A., Menon, V.: *Jumbo Express- A scientific approach to understanding and mitigating elephant mortality due to train accidents*. WTI (2001)
20. Athreya, V.R., Balsare, A.V.: *Human-leopard conflict management guidelines* (2007)
21. Joshi, R.: *Train accidental Deaths Of LeopardsPanthera Pardus in Rajaji National Park: A Population in Threat*. *World Journal Of Zoology* 5(3), 156–161 (2010)
22. De Man, A.P.: *Dynatrack, a survey of dynamic railway track properties and their quality*, Dissertation TU Delft, DUP-Science (December 2002) ISBN 9-406-2355-9

An Iterative Suffix Stripping Tamil Stemmer

Vivek Anandan Ramachandran and Ilango Krishnamurthi

Department of Computer Science and Engineering,
Sri Krishna College of Engineering and Technology,
Coimbatore – 641 008, Tamilnadu, India
{rvivekanandan, ilango.krishnamurthi}@gmail.com

Abstract. Stemming algorithm is a procedure that attempts to map all the derived forms of a word to a single root, the stem. It is widely used in Information Retrieval applications with the main objective of enhancing the recall factor. Stemmer is also used in various other applications such as summarization, classification and text mining etc. Apart from English, exploration on developing stemmers for both the native and the regional languages are also being carried out. In this paper, we present a stemmer for Tamil, a Dravidian language. Our stemmer effectiveness is 84.79%.

Keywords: Information Retrieval, Stemming, Tamil.

1 Introduction

Stemming is the process of mapping all the derived forms of a word to a single root, the stem. In stemming, it is not necessary for the resultant to be a genuine linguistic word. For example, applying stemming to the words *operating*, *operates* and *operator* maps each one of them to the stem *oper*. Here the resultant *oper* is not a proper linguistic word.

Stemming is used in many Information Retrieval (IR) applications to improve the recall factor [1]. This is due to the reason that a word in its derived forms tends to have the similar meaning. For example, when a user wants to search the documents with the term *operating* they might also need the documents with the term *operates* or *operated*. Further, as many forms of a single word are mapped to a single entity, stemming assists in decreasing the size of an index file used in an IR system.

Due to the emergence of Cross-Lingual IR (CLIR), exploration on developing stemmers for both the native and the regional languages are carried out. Despite many studies, there is no readily available stemming algorithm for Tamil¹[2]. In this paper, we propose a suffix stripping stemmer for Tamil.

This paper is organized as follows: In Section 2 we discuss the researches related to our stemmer. In Section 3 we brief about the Tamil suffixes. The difficulties in

¹ Tamil is a Dravidian language spoken predominantly by the Tamil people of the Southern India. It has official status in India, Sri Lanka and Singapore. Substantial minorities in Malaysia, Mauritius and Vietnam also speak Tamil. Throughout the paper transliterated form of Tamil as quoted in the Appendix Section is used.

developing stemmer for Tamil language are highlighted in Section 4. In Section 5 we describe about our algorithm. In Sections 6 and 7, we present the evaluation analysis and the concluding remarks respectively.

2 Related Work

Many exhaustive studies to develop stemmers are being done since 1960s[3][4]. Because of such investigations, good stemmers have emerged. However, those investigations were majorly dealt in English. Because of the egression of CLIR in 1990s, there was an increased demand from the research community to develop stemmers for both the native and the regional languages. Due to this demand, stemmers were devised for many languages.

In the case of Indian languages, stemming was first reported for Hindi in 2003[5]. Slowly investigations for other languages such as Bengali[6], Urdu[7], Malayalam[8] and Punjabi [9] were also carried out. However, there is no readily available stemmer for Tamil. In this paper, we present our research experiences in developing a Tamil Stemmer.

To develop a stemmer for Tamil or any other languages, a basic approach to carry out the process is required. The most common approaches used for developing a stemmer are Brute force, Affix Stripping, N-Gram, Hidden Markov Model (HMM), Corpus based technique, Clustering method, Finite-State-Automata method, Morphological process, String distance measure, Hybrid approaches. Among all the existing approaches, we make use of affix stripping because of its inherent support to develop a stemming algorithm in an easier and faster way.

Most of the existing stemmers remove the suffixes based on the longest matching word. For example among the matching suffixes *ates*, *tes*, *es* and *s* existing in the word operates, the suffix *ates* will be removed by a stemmer. Similar to most of the existing stemmers, we remove the suffixes based on the longest matching word.

To develop a stemmer for a language, a preliminary study on the possible suffixes of a word in the corresponding language need to be taken. In the next Section, we explain about the possible suffixes for Tamil.

3 Tamil Suffixes

Developing a stemmer for any language needs a study on the derived forms that a word in the corresponding language can have. Tamil is a highly inflected language. In Tamil, a word will contain a root to which one or more affixes are attached. The affixes can be either a prefix or a suffix. In most of the cases, Tamil affixes are suffixes. There is no exact limit for attaching the number of suffixes to a word.

To know about the possible suffixes that a word in Tamil language one can refer to the flow charts [10] and [11]. Besides considering the possible suffixes forms of a word, a stemming algorithm has to consider certain standard computing issues. In the next Section, we explain the computing issues considered for developing our stemmer.

4 Computing Issues

Developing a Tamil stemmer is not a straightforward task. In this Section, we explain the major difficulties faced by us while designing the algorithm. They are briefed as follows:

4.1 Homographs

Homographs are the words that have identical pronunciations but different meanings. In Tamil, there are numerous instances of homographs. For example, the word *aaNTavan* denotes either the noun *God* or the verb *ruled by a male* formed from the root *aaL*. It is difficult for a rule-based stemmer to map such terms to their root. Hence, we decided to frame our algorithm by do not considering the homograph issues.

4.2 Irregular Verbs

Irregular verbs do not follow standard patterns in their tense form. For example, the past tense of the verb *say* is *said* and not *sayed*. Mapping such forms of word to a single root is a difficult task. Such cases exist in Tamil also. For example, the past, present and future tense of the verb *sol* (*say*) are *sonneen* (*I said*), *solkiReen* (*I am saying*) and *solveen* (*I will say*) respectively. Following the standard patterns the past tense for *sol* should be *solneen*. However, this is not correct. Devising rules to handle such case is arduous as it needs a deep look up dictionary. Hence, we decided to consider this issue in the future version.

4.3 Proper Noun Derivations

A proper noun usually indicates a particular thing. In certain cases, proper noun end letters match with the normal suffixes. Assuming those patterns to be suffixes, most of the stemmers remove them from the proper noun. For example, Porter stemmer maps the proper noun *operator* to *oper* due to the assumption that *ator* is a suffix. To overcome such cases, it is very difficult to realize a proper noun by framing hand-crafted rules. Therefore, similar to most of the algorithms, if the common suffix patterns exist in a proper noun we decide to stem it. For example, our approach maps *iyakkiyavan* (*A person who operated*) to *iyakki* (*operated*).

4.4 Handling Non Derived Words Ending with Usual Suffix Pattern

In some cases, word ends with few patterns that match with a suffix but that pattern does not denote a suffix. For example:

- In English, the word *sing* contains *ing* which usually denotes a suffix but not in this case.
- In Tamil, the word *vitai* (*seed*) contains *ai* which usually denotes a suffix.

It could be inferred that to handle this case a stemmer needs a heavy look-up table and this table cannot be constructed easily. So we decided to handle this case in the future versions.

4.5 Handling Agglutinative Case

Tamil is an agglutinative language; a compound word can be formed from two or more simple words without changing the meaning of the simple words. For example consider the word *maJainiir* (Rain Water) formed from two simple words *maJai* (Rain) and *niir* (Water). Consider the word *maJainiiri_n* (of the rain water). The word is a derived form of *maJai*. Mapping the word *maJainiiri_n* to the simple word *maJai* is a laborious task. So we decided to neglect mapping compound word to simple word.

Apart from the above discussed computational issues, proper computational steps should also be designed to develop a good stemmer. In the next Section, we explain the design portion of our stemmers.

5 Design

Table 1. Stemming Algorithm

Input	Tamil String (<i>Input</i>), Suffix List (<i>SL</i>)
Output	Root of the <i>Input</i> (<i>Output</i>)
Prerequisite	The <i>SL</i> should be stored in descending order of suffixes length
<pre> Function String stem (<i>Input</i>) Begin 1. String <i>Output</i>, 2. String <i>Temp-Output</i>= ruleBase(<i>Input</i>) 3. While <i>Input</i> != <i>Temp-Output</i> a. <i>Temp-Output</i> = ruleBase(<i>Input</i>) b. <i>Input</i> = <i>Temp-Output</i> 4. return <i>Output</i> End Function String ruleBase (<i>temp</i>) Begin 1. Flag = true 2. While (Flag) a. Iterate all the suffix one by one i. If the <i>temp</i> ends with any suffix (say <i>Sf</i>) A. <i>temp</i> = <i>temp</i> - <i>Sf</i> b. return <i>temp</i> End </pre>	

Generally, for removing multiple suffixes existing in a word of any language iterative stemmer is used. An iterative stemmer starting from the end of the inflected input word will remove a longest matching suffix at a time and progress towards the root. The algorithm pseudo-code is presented in Table 1. The list of suffixes that our stemmer can handle is listed in Table 2.

Table 2. Tamil Suffixes

<i>etirttaaRpool</i>	<i>appuRam</i>	<i>tavira</i>	<i>teRku</i>	<i>uTa_n</i>	<i>oTTi</i>	<i>kiR</i>
<i>aTuttaaRpool</i>	<i>koNTiru</i>	<i>muulam</i>	<i>aa_na</i>	<i>iyal</i>	<i>iTam</i>	<i>een</i>
<i>veeNTiyiru</i>	<i>veeNTum</i>	<i>piRaku</i>	<i>tolai</i>	<i>avaL</i>	<i>kiiJ</i>	<i>aaL</i>
<i>uNkaLee_n</i>	<i>etirkku</i>	<i>pi_npu</i>	<i>ava_n</i>	<i>mu_n</i>	<i>ttal</i>	<i>uL</i>
<i>vaJiyaaka</i>	<i>aayiRRu</i>	<i>pakkam</i>	<i>illai</i>	<i>avai</i>	<i>tiir</i>	<i>um</i>
<i>varaikkum</i>	<i>veLiyil</i>	<i>umee_n</i>	<i>poola</i>	<i>avar</i>	<i>paar</i>	<i>tt</i>
<i>veeNTivaa</i>	<i>kuRittu</i>	<i>meelee</i>	<i>aakum</i>	<i>a_na</i>	<i>atu</i>	<i>il</i>
<i>mu_n_naal</i>	<i>maatiri</i>	<i>kki_nR</i>	<i>kiTTa</i>	<i>paTi</i>	<i>aam</i>	<i>pp</i>
<i>patilaaka</i>	<i>paarttu</i>	<i>taaNTi</i>	<i>ki_nR</i>	<i>viTa</i>	<i>aar</i>	<i>ai</i>
<i>tavirttu</i>	<i>naTuvil</i>	<i>uTaiya</i>	<i>pooTu</i>	<i>meel</i>	<i>aay</i>	<i>al</i>
<i>illaamal</i>	<i>vaTakku</i>	<i>etiree</i>	<i>oJiya</i>	<i>kiJi</i>	<i>kka</i>	<i>ya</i>
<i>veeNTaam</i>	<i>meeRku</i>	<i>uNkaL</i>	<i>paNNu</i>	<i>ukku</i>	<i>iir</i>	<i>nt</i>
<i>kuRukkee</i>	<i>kuuTum</i>	<i>aarkaL</i>	<i>koNTu</i>	<i>viTu</i>	<i>kaL</i>	<i>a</i>
<i>allaamal</i>	<i>aTiyil</i>	<i>vaittu</i>	<i>aNTai</i>	<i>chey</i>	<i>oom</i>	<i>p</i>
<i>varaiyil</i>	<i>etiril</i>	<i>kiiJee</i>	<i>uLLee</i>	<i>kiTa</i>	<i>poo</i>	<i>t</i>
<i>kuuTaatu</i>	<i>aaTTam</i>	<i>arukee</i>	<i>taLLu</i>	<i>muTi</i>	<i>vaa</i>	<i>u</i>
<i>veLiye</i>	<i>appaal</i>	<i>iirkaL</i>	<i>paRRi</i>	<i>ooTu</i>	<i>i_n</i>	<i>v</i>
<i>iTaiyil</i>	<i>iIruntu</i>	<i>kaaTTu</i>	<i>pinti</i>	<i>koTu</i>	<i>tal</i>	
<i>aakaatu</i>	<i>chuRRi</i>	<i>nookki</i>	<i>patil</i>	<i>kkiR</i>	<i>vai</i>	
<i>kiJakku</i>	<i>arukil</i>	<i>viTTu</i>	<i>mutal</i>	<i>aa_n</i>	<i>iru</i>	

Consider the input derived word *paRkaLi_nil* (in the teeth) derived from the word *pal(teeth)*. During the first iteration suffixes *il* and *i_n* will be removed. During the second iteration *kaL* will be removed respectively. The output will be *paR*. Although the algorithm for stemming Tamil words is designed successfully, it has to be evaluated. In the following Section, we present the analysis carried out by us to study the algorithm's effectiveness.

6 Evaluation

In general, evaluation signifies the act of assessing something. To evaluate our stemmer we implemented our algorithm in Java. A sample screen shot of our stemmer is shown in Figure 1.

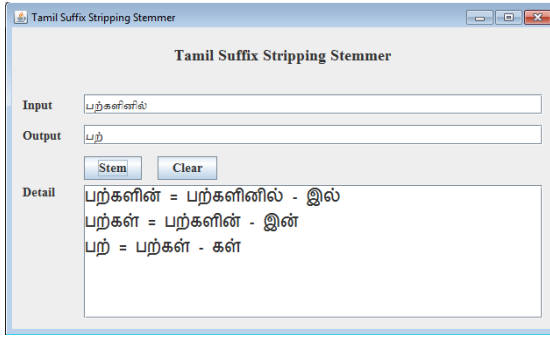


Fig. 1. Sample screen shot of the Tamil Suffix Stripping Stemmer

After developing any stemmer it is important to analyze its capability, i.e., to assess the range of the words that the system is able to stem properly. We requested two students, none of whom are directly or indirectly involved with our project to generate the corpus. They developed a Tamil corpus containing 10059 words derived from 783 roots. The corpus is framed from different portions of Tamil newspapers confining to various domains such as Business, Classifieds, Entertainment, Politics and Sport.

It is important to remember as stated in the Introduction that the output of a stemmer need not be a proper linguistic word. Therefore, correctness of a stemmer does not denote the linguistic correctness. A stemmer is said to be accurate if it conforms to the following conditions:

- If it maps all the derived forms of a word to a single root.
- The words mapped by it to a single stem are genuine linguistic variants.
- If it does not stem a non-suffix from a word.

If a stemmer does not map all the considered derived forms of word to a single stem then the phenomenon is called *Understemming*. An instance of *Understemming* is a stemmer conflating *tried* to *tri* and *try* to *try* instead of conflating both to *try*. If a stemmer conflates the words to a single stem that are genuinely linguistic invariants then the phenomenon is called *Overstemming*. An instance of *Overstemming* is a stemmer conflating both the words *cares* and *cars* to *car*, instead of conflating *cares* to *care* and *cars* to *car*. If a stemmer removes a nonsuffix from a word, it is called *Mis-stemming*. For example, conflating the words *reply* to *rep* instead of conflating it to *reply* is called *Mis-stemming*. Most of the stemmers do not give importance to *Mis-stemming*. This is because it does not spoil the recall factor in an IR application.

Due to the same reason, we evaluate our stemmer using only *Understemming* and *Overstemming*. They are calculated using the following formulas (1) and (2) respectively.

$$\text{Understemming} = (\text{Number of variants understemmed}/\text{Total variants}) * 100\% \quad (1)$$

$$\text{Overstemming} = (\text{Number of variants overstemmed}/\text{Total variants}) * 100\% \quad (2)$$

Evaluating our approach using the above-mentioned corpus containing 783 root variants, we found that 32 and 87 were *understemmed* and *overstemmed* respectively. Hence, The *Understemming* and *overstemming* values are 4.09 % and 11.12 % respectively. Our stemmer effectiveness is calculated using the formula (3).

$$\text{Stemmer Effectiveness} = 100\% - [\text{Overstemming \%} + \text{Understemming \%}] \quad (3)$$

Effectiveness for our approach is 84.79 %. This is considerably a good value. Yet the reason behind achieving a moderate effectiveness value is due to the factors discussed in the Section 4.

7 Conclusion

In this paper, we hypothesize an Iterative Tamil Stemmer. This paper demonstrates preliminary explorations. Yet there is a lot to be achieved. We will be glad if anyone could provide valuable suggestions or selectively enhance our work. In addition, it will be encouraging if someone can suggest suffixes to be included or excluded from our rule-base to enhance the recall factor.

References

1. Kraaij, W., Pohlman, R.: Viewing Stemming as Recall Enhancement. In: The Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 40–48 (1996)
2. Germann, U.: Building a Statistical Machine Translation System from Scratch: How Much Bang for the Buck Can We Expect? In: ACL 2001 Workshop on Data-Driven Machine Translation, Toulouse, France, July 7 (2001)
3. Lovins, J.B.: Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics* 11(1), 22–31 (1968)
4. Porter, M.F.: An Algorithm for Suffix Stripping. *Program* 14(3), 130–137 (1980)
5. Ramanathan, A., Rao, D.: A lightweight stemmer for Hindi. In: The Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL) for South Asian Languages Workshop (April 2003)
6. Islam, Z., Uddin, N., Khan, M.: A Light Weight Stemmer for Bengali and Its Use in Spelling Checker. In: Proceedings of 1st International conference on Digital Communications and Computer Applications (DCCA 2007), Irbid, Jordan, pp. 87–93 (2007)
7. Akram, Q.-U.-A., Naseer, A., Hussain, S.: Assas-Band, an affix-exception-list based Urdu stemmer. In: Proceedings of the 7th Workshop on Asian Language Resources (2009)

8. Malayalam Stemmer,
http://nlp.au-kbc.org/Malayalam_Stemmer_Final.pdf
9. Kumar, D., Rana, P.: Design and Development of a Stemmer for Punjabi. International Journal of Computer Applications 11(12), 0975–8887 (2010)
10. Tamil Noun Flow Chart, http://www.au-kbc.org/research_areas/nlp/projects/morph/NounFlowChart.pdf
11. Tamil Verb Flow Chart, http://www.au-kbc.org/research_areas/nlp/projects/morph/VerbFlowChart.pdf

Appendix: Tamil Transliteration Scheme Used in This Paper

A	Aa	i	ii	u	uu	e	ee	ai	o	oo	au	q					
அ	ஆ	இ	ஈ	உ	ஊ	எ	ஏ	ஐ	ஓ	ஔ	ஔள	ஔி					
k	-N	ch	-n	_n	T	N	t	n	p	m	y	r	l	v	J	L	R
க்	ங்	ச்	ஞ்	ன்	ட்	ண்	த்	ந்	ப்	ம்	ய்	ர்	ல்	வ்	ழ்	ள்	ற்

Combining Classification Algorithm with DOM Algorithm for Web Information Extraction – A Hybrid Approach

Venkat Ramana Bhavanasi¹ and A. Damodaram²

¹ Dept.of CSE MIPGS, Hyderabad, India
JNIAS-JNTUH

² Dept. of CSE, JNTUH, Hyderabad
venkatbhavanasi@gmail.com,
damodarama@jntu.ac.in

Abstract. our approach is to merge information extraction algorithm for Web sources with a classification algorithm, by doing this the algorithm overcomes the shortcoming of DOM-based information extraction which is not enough to adapt to the structural change of information in Web pages. In addition, a detection strategy is used to extract Web pages with multiple records. According to the change of similarity matrixes, it can discover the dividing point of records and then extracts all the records. Experiments show that when it is trained by a single sample Web page, the algorithm can still obtain a good result in precision and recall.

Keywords: Information Extraction, Classification, DOM algorithms, Web Mining, Deep web.

1 Introduction

The explosive growth and popularity of the world-wide web has resulted in a huge amount of information sources on the Internet. However, due to the heterogeneity and the lack of structure of Web information sources, access to this huge collection of information has been limited to browsing and searching. Sophisticated Web mining applications, such as comparison shopping robots, require expensive maintenance to deal with different data formats. To automate the translation of input pages into structured data, a lot of efforts have been devoted in the area of Information Extraction (IE). IE produces structured data ready for post-processing, which is crucial to many applications of Web mining and searching tools [1].

Web IE is implemented by wrappers. The traditional wrapper generation depends on manual code which is waste of time and easy to make a mistake. In order to solve the problem, many methods are presented to generate wrappers automatically and semi-automatically. IE based on DOM mainly [2, 3] has poor adaptability to the structural change of information in Web pages such as losing items. In order to improve precision and recall, [4, 5] are based on XML technology to extract information from Web pages. It also combines XSLT and XPath technologies.

Because the main extraction rule is the path of information in XML, the adaptability to the structural change of information is still not perfect. [6] Proposes a method that can extract information from different sites automatically based on keywords of a certain topic and the distance of nodes. This method needs to establish a keyword library firstly. It is not easy and leads heavy workload. [7] Obtains effective information by the start and end anchor point of HTML tags. Because of the similarity of HTML tags, the extraction process could be disturbed by the data that has similar structure.

2 Description of Extraction Processes

The extraction process can be made of two parts. One refers to the processing of the sample Web page and the other refers to the practical Web page

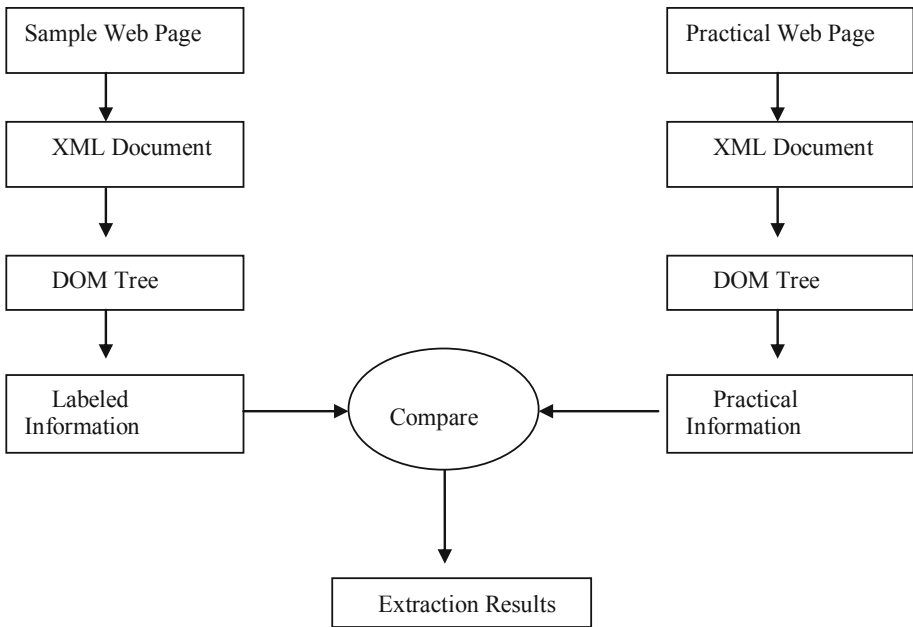


Fig. 1. The Extraction Process involving Sample and Practical Webpage [1]

The sample Web page is transformed into a well-formed XML document first. At the same time some useless nodes are removed such as span, script and so on. Then the system converts the XML document into a DOM tree. The user can label information in the DOM tree and input previous character strings which are relevant to the information of interest. The system records the input of the user and saves the characteristics of labeled information. After processing the sample-Web page, a practical-Web page which will be extracted is inputted. The system cleans the Web

page and converts it into a DOM tree. Then it obtains the characteristics of all the nodes in the DOM tree and saves them. The characteristics of information the user labeled and those from the practical Web page are compared at last. The system finds the most similar information in the practical Web page as the extraction result.

The Web pages are standardized and converted into DOM trees according to [8]. Then we transform the Web pages into well-formed XML documents by Tidy and use JTree to show the DOM tree of the sample Web page. Finally, the information can be labeled.

The Algorithms based on DOM or XML, using single path characteristic as an extraction rule usually can't get an ideal result. A tiny structural change of information can cause the extraction rules invalidation. So according to the feature of a DOM tree, we add some other characteristics, including the name of parent node, the attribute, the name of left neighboring node and the name of right neighboring node. Moreover similar information between the sample Web page and the practical Web page usually has the same previous character string.

When the document is parsed into a DOM tree, it is easy to get some characteristics of the current information node such as the name of parent node, the attribute, the name of left neighboring node and the name of right neighboring node according to the methods that DOM provides. The acquirement of path characteristic as well as the relative distance from the previous character string to its corresponding information node is a little complex. The path characteristic is an expression that is formed by HTML tags and indexes. A circulation algorithm from the current node to its parent node is used to obtain the path characteristic of the information node. After the path characteristic in the sample Web page is obtained, the corresponding number will be checked. If it is "Multiple", in order to obtain a wide coverage extraction rule, the last HTML index of path characteristic will be removed. The residual part is considered as a new path characteristic instead of the original one. If some information nodes have the same previous character string, the common path will be their new path characteristic.

The essence of this method is to find needful information in the practical Web page by characteristic comparison. Information nodes whose characteristics are similar to those the user labeled in the sample Web page are found. The specific comparison is shown as follows. We divide the path characteristic into two parts [9]. One refers to HTML tags sequence, the other refers to HTML indexes sequence. Let PA be the path characteristic in the sample Web page. Its tags sequence is defined as $PMA = (aM1; aM2; aM3; :::; aMn)$. Its HTML indexes sequence is defined as $PNA = (aN1; aN2; aN3; :::; aNn)$. Let PB be the path characteristic in the practical Web page. Its tags sequence is defined as $PMB = (bM1;bM2;bM3; :::;bMn)$. The HTML indexes sequence is defined as $PNB = (bN1; bN2; bN3; :::; bNn)$. In a DOM tree the closer HTML tags and indexes are to the root node, the greater influence they have on the similarity of information nodes. According to the difference of position different weights should be given when calculating the similarity [10].

$$PMS(PMA, PMB) = C. \sum_{i=1}^n 2^{-(i-1)} \cdot [(aMi = bMi)? 1: 0] \tag{1}$$

Considering HTML indexes, the more similar two information nodes are, the smaller the difference of the indexes is in the relevant position. So the similarity between *PNA* and *PNB* is as follows.

$$PNS(PNA, PNB) = C \cdot \sum_{i=1}^n 2^{-(i-1)} \cdot (|a N_i - b N_i| + 1)^{-1} \tag{2}$$

C is the normalization factor above two formulas and its value is $1 / \sum_{i=0}^{n-1} 2^{-i}$

Given equal weights to two characteristic components, the similarity between *PA* and *PB* is calculated as follows.

$$PS(PA; PB) = 1/2 \cdot PMS + 1/2 \cdot PNS \tag{3}$$

In the practical Web page, if the information node contains corresponding previous character string in the relative distance of the sample Web page, *FS* (*FA*; *FB*) is 1, otherwise it is 0. The similarity of the name of parent node is

$$PaS(PaA, PaB) = (PaA = PaB) ? 1 : 0 \tag{4}$$

The similarity of the attribute is

$$AtS(AtA, AtB) = (AtA = AtB) ? 1 : 0 \tag{5}$$

The formulas above mean that if the characteristic of information node in the sample Web page is the same to that in the practical Web page, the expression takes 1 as its value, otherwise it takes 0. Next the total similarity *Sim* between information *A* in the sample Web page and *B* in the practical Web page is calculated.

$$Sim(A;B) = 1 / 7 \cdot (PS + 2 \cdot FS + PaS + AtS + LS + RS) \tag{6}$$

The information in the practical Web page which has the maximum similarity with that in the sample page is saved. By now the extraction process has been completed. According to some Web pages with losing items, a Lowest Similarity Degree (LSD) is used to avoid extracting error objects.

3 Extracting Web Pages with Multiple Records

Because the training is based on a single sample, the precision and recall are not satisfactory when extracting multi-record Web pages especially the Web pages with indefinite records. Here a detection strategy is used to solve this problem 1. The user labels each record in the sample Web page first. Then all the records' common path which is defined as *P* is obtained. After that, we compare the path of first record with the common path. The record's first index which appears after the common path is defined as *i*. If all the records have the same information nodes, the system will save the first record. If there are some records with losing items, the record with the most information nodes will be saved.

4 Experiment Results

To evaluate the extraction result we use precision and recall. The data source used in the experiment comes from IMDB. A Web page is selected randomly from the Web pages in IMDB and annotated as the sample page. Then we extract another 50 Web pages. To avoid any accidental occurrence, tests are repeated 3 times with different sample pages. The average precision and recall are taken. The comparison has been made with [5] and [7] under the same condition. Table 1 gives the final result.

Table 1. Extraction results from IMDB

Algorithm	Total Pages	Precision (%)	Recall (%)
Dong's Algorithm	75	91.2	92.4
Yong's Algorithm	75	94.5	90.3
Ours Algorithm	75	99.6	97.8

Result in Table 1 shows that the improved DOM-based method can obtain a higher precision and recall compared to method 1 and method 2. Even if some Web pages lose items occasionally, it can still get a better extraction result.

In order to validate the effectiveness of the detection strategy, another data source OKRA is used to carry on the experiment. Web pages in OKRA are multi-record pages. We select one page as the sample page and extract another 50 pages. The comparison with [5] and [7] is given in Table 2.

Table 2. Total Extraction Results from OKRA

Algorithm	Total Pages	Precision (%)	Recall (%)
Dong's Algorithm	75	89.2	89.3
Yong's Algorithm	75	100	100
Ours Algorithm	75	100	100

5 Conclusions

This paper proposes an improved DOM-based method for Web information extraction. The extraction process is implemented by comparison and classification. When extracting multi-record Web pages, a detection strategy is used to determine the dividing point between records. Then the records can be obtained one by one. It reduces the difficulty of extraction. Experiments show that the method has a better performance with IMDB and OKRA.

References

1. Zhang, L., Li, M., Dong, N., Wang, Y.: An Improved DOM-based Algorithm for Web Information Extraction. *Journal of Information & Computational Science* 8(7), 1113–1121 (2011)

2. Dewanto, S., Mustofa, K.: IE using Automatic pattern discovery based on tree matching computer Science Department, GM University, pp. 342–348 (2011)
3. Franck, R., Olivier, T., Ronan, T.: Finding an application-appropriate model for XML data warehouses. *Information Systems* 35(6), 662–687 (2010)
4. Dong, M., Fang, S., Yang, Z.P.: Jtree and XPath based information extraction on dynamic web pages. *Journal of Information*, 73–75 (2007)
5. Deng, J.S., Zheng, Q.L., Peng, H., Lin, X.D.: Web pages information retrieval based on keywords cluster and node instance. *Computer Science* 34(4), 213–216 (2007)
6. Yeonjung, K., Jeahyun, P., Taehwan, K., dan Joongmin, C.: Web IE by HTML Tree Edit Distance Matching. In: *Proceedings of the International Conference on Convergence Information Technology (ICCIT 2007)*, Washington, DC, US (2007)
7. Zhai, Y., Liu, B.: Structured Data Extraction from the Web Based on Partial Tree Alignmen. *IEEE Transaction on Knowledge and Data Engineering* 16(12), 1614–1628 (2006)
8. Hedley, Y.L., Younas, M., James, A., Sanderson, M.: Sampling, information extraction and summarization of hidden web databases. *Data and Knowledge Engineering* 59(2), 213–230 (2006)
9. Liu, H.: Eb information extraction and corpus construction system with C sharp. *Computer Engineering* 32(16), 49–51 (2006)
10. Chang, C.H., Kayed, M., Girgis, M.R., et al.: A survey of web ie systems. *IEEE Transactions on Knowledge and Data Engineering* 18(10), 1411–1428 (2006)
11. Li, X.Y., Zhang, Y.F., Lu, J.J., Xu, B.W.: A classification method for web information extraction. *Wuhan University Journal of Natural Sciences* 9(5), 823–827 (2004)
12. Yang, W.Z., Xu, L.H., Chen, S.F., et al.: Design and implementation of XPath-based web information extraction. *Computer Engineering* 29(16), 82–83 (2003)
13. Yu, K., Cai, Z., Mi, Z.C., Cai, Q.S.: Information retrieval method based on path learning. *Mini-micro Systems* 24(12), 2147–2149 (2003)
14. Li, X.D., Gu, Y.Q.: DOM-based information extraction for the web sources. *Chinese Journal of Computers* 25(5), 526–533 (2002)
15. The dataset resource, <http://www.imdb.com/chart/top?tt00668646>
16. The dataset resource, http://www.isti.edu/info_agents/rise/repository.html

Computerized Lesion Detection in Colposcopy Cervix Images Based on Statistical Features Using Bayes Classifier

Pazhayanoor Seethapathy RamaPraba¹ and H. Ranganathan²

¹ Sathyabama University,
Chennai, India

pazrama@yahoo.co.in

² Sakthi Mariamman Engineering College,
Chennai, India

Abstract. Colposcopy is one of the methods for cervical cancer screening that uses visual testing based on the color change of abnormal cells that turn white when exposed to acetic acid called AcetoWhite (AW) region. In this paper, a novel approach to detect the AW region in the cervix image based on statistical features and Bayes classifier is presented. Colposcopic images are acquired in raw form, contains cervix, regions outside the cervix and parts of the imaging devices. In the preprocessing stage the irrelevant information in the cervical images are removed based on Mathematical morphology and Gaussian Mixture Modeling and also specularities are removed based on HSI colour space. The detection of lesion (AW region) is achieved by extracting the statistical features such as mean, standard deviation and skewness and the features are used as an input to the Bayes classifier. Segmentation results are evaluated on 260 images of colposcopy.

Keywords: Colposcopy, Cervigram, Gaussian Mixture Model, Fuzzy C means Clustering, Morphological Operations, Statistical Features.

1 Introduction

Cervical cancer is the most common form of cancer in women under 35 years of age, worldwide. Cervical cancer is largely preventable and curable with regular pap tests and pelvic exams. Pap test is used to find cell changes in the cervix. In a Pap test, the nurse or the Doctor uses a speculum to view the vagina and takes a few cells from the cervix using a soft brush. A lab then checks these cells for the presence of cancer. As per the latest news reported in the Hindu dated July 9, 2011, Colposcopy helped early detection of Cervical Cancer in women. Colposcopy is a medical diagnostic procedure to examine the illuminated, magnified view of the cervix and the tissues of vagina and vulva. Colposcopic images are characterized by color, texture and relief information. Thus, their automatic analysis is difficult. However, the diagnosis of experts about some much debated images is often different, because of the very high specialization required. An integrated analysis tool for helping gynecologists to build their

colposcopic diagnosis is proposed by Isabelle Claude and Philippe Pouletaut [1]. Moreover, specific preprocessing methods and different segmentation methods are available like principal component analysis and multidimensional histogram analysis.

Cerviographic images are known as cervigrams. In cervigram, the lesions are of varying sizes and complex, non-convex shapes. A new methodology that enables the segmentation of non-convex regions, thus providing a major step forward towards cervigram tissue detection and lesion description is presented by Shiri Gordon and Hayit Greenspan [2]. The framework transitions from pixels to a set of small coherent regions, which are grouped bottom-up into larger, non-convex, perceptually similar regions, utilizing a new graph-cut criterion and agglomerative clustering. A multistage scheme for segmenting and labeling regions of anatomical interest within the cervigrams is presented by Hayit Greenspan and Shiri Gordon [3]. In particular, focusing on the extraction of the cervix region and fine detection of the cervix boundary, specular reflection is eliminated as an important preprocessing step and in addition, the entrance to the endocervical canal is detected.

Colposcopic image classification based on contour parameters using different artificial neural network and the KNN classifier is proposed by Claude and I. Winzenrieth [4]. A set of original spatial and frequency parameters is extracted from 283 samples to characterize the attribute of contour. The spatial parameter is the number of the region around the edges and the frequency parameters are amplitude of first peak, frequency of the end of first peak, area under first peak and area under other peaks. Then the Principal Component Analysis is performed to test the parameters. Segmentation and classification of cervix lesions by pattern and texture analysis is presented by Bhakti Tulpule and Shuyu Yang [5]. The acetowhite region, a major indicator of abnormality in the cervix image, is first segmented by using a non-convex optimization approach. Within the acetowhite region, other abnormal features such as the mosaic patterns are then automatically classified from non-mosaic regions by texture analysis.

A cost-sensitive 2v-SVM classification scheme to cervical cancer images to separate diseased regions from healthy tissue is proposed by Yusuf Artan and Xiaolei Huang [6]. Multiplier classifier scheme is used instead of the traditional single classifier to test the NCI/NLM archive of 60000 images. The phase correlation method followed by a locally applied algorithm based on the normalized cross-correlation is presented for image registration by Acosta-Mesa Héctor G and Zitová Barbara [7]. During the parameterization process, each time series obtained from the image sequences is represented as a parabola in a parameter space. A supervised Bayesian learning approach is proposed to classify the features in the parameter space according to the classification made by the colposcopist.

Cervical Intraepithelial Neoplasia (CIN) is detectable and treatable precursor pathology of cancer of the uterine cervix. A non-parametric technique, based on the transformation and analysis of the distortion-rate curve is proposed by Yeshwanth Srinivasan and Enrique Corona [8] to assess the model order. This technique provides good starting points to infer the GMM parameters via the expectation-maximization (EM) algorithm, reducing the segmentation time and the chances of getting trapped in local optima.

2 Proposed Method

A novel computerized system approach proposed in this paper can be used for diagnosing the (AW) lesion in the cervix image. The process is carried out in three parts, namely, ROI segmentation of cervix, Removal of specularities and Detection of AW Lesion.

2.1 ROI Segmentation of Cervical Images

The steps in segmentation of cervix from the colposcopy cervix image are shown in Figure 1. An ROI segmentation image processing system substantially masks non-ROI image data from a digital image to produce a ROI segmented image for subsequent digital processing. A colposcopy cervical image contains major cervix lesions, regions outside the cervix and parts of the imaging device. In this method, only the major cervix lesion is segmented for further processing. The major cervix lesion is a reddish, nearly circular section approximately centered in the image. This feature is used to identify the ROI region.

For ROI segmentation, first the given cervix image in RGB colour space is converted into Lab colour space due to the fact that Lab colour space is a good choice for representing the colour. The Euclidean distance of a pixel from the image center is extracted for all pixels and it is represented as Euclidean distance array d . The Gaussian Mixture Model (GMM) parameters μ and σ are calculated by (1) for the Euclidean distance array and the colour channel a from the Lab colour space.

$$a(x, y) = \frac{a - \mu_a}{\sigma_a} \quad \text{and} \quad d(x, y) = \frac{d - \mu_d}{\sigma_d}. \quad (1)$$

By using the GMM parameters, Euclidean distance array and the colour channel a from the Lab colour space are normalized and aggregated into a single array which is given to Fuzzy C means clustering (FCM) algorithm as an input. In FCM, the total number of cluster is set to 2. Among the 2 cluster, the cluster which has the smallest d and largest a is chosen as ROI after the cluster centroids are de-normalized. Finally, morphological opening is used to remove the small regions and fill the holes to get the required ROI image that contains only the cervix lesion.

2.2 Specular Reflection of Cervical Images

Specular reflections (SR) appear as bright spots heavily saturated with white light. These occur due to the presence of moisture on the uneven cervix surface, which acts like mirrors reflecting the light from the illumination source. The block diagram for removal of specular reflection in the segmented cervix image is shown in Figure 2.

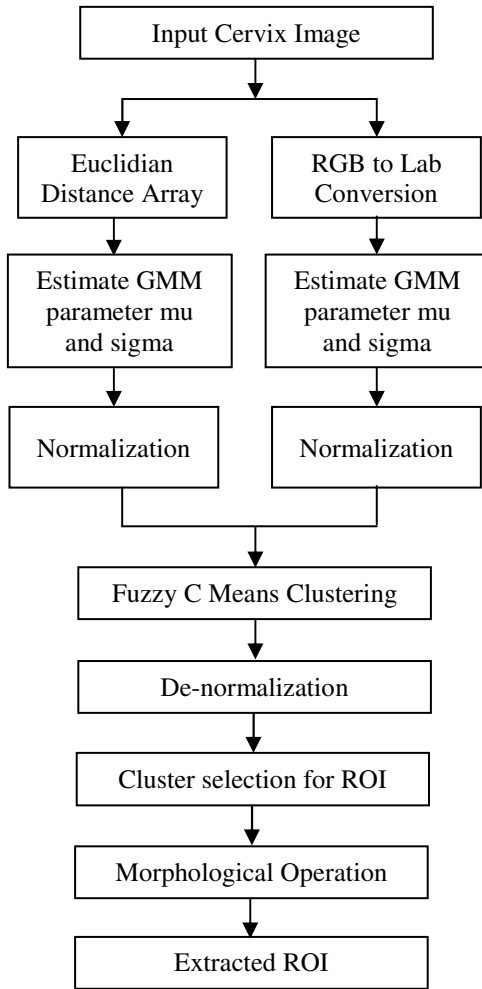


Fig. 1. Block diagram of ROI Segmentation of cervical images

The segmented cervix image in RGB colour space is converted into *HSI* colour space due to the fact that *HSI* colour space represents the colour similarity how the human eye senses colours. The *HSI* color model represents every color with three components, hue (*H*), saturation (*S*) and intensity (*I*). Specularities always have very intense brightness and low saturation values. Hence, the *I* and *S* component in the *HSI* colour space is used to find the SR region and the conversion formulae are given in (2) and (3) respectively.

$$I = \frac{R + G + B}{3} . \tag{2}$$

$$S = 1 - \frac{\min(R, G, B)}{3} . \tag{3}$$

The initial SR regions are identified by applying thresholding technique on image pixels. The threshold values are defined in (4)

$$I > 0.8 * I_{max} \quad S > 0.6 * S_{max} . \quad (4)$$

After thresholding, morphological dilation is performed on the thresholded image by using square structuring element of width 5 to get SR regions. Boundaries are extracted from the SR regions by using 8-connected neighborhood. Finally, the SR regions are smoothly interpolates inward from the pixel values on the boundaries by solving Laplace's equation.

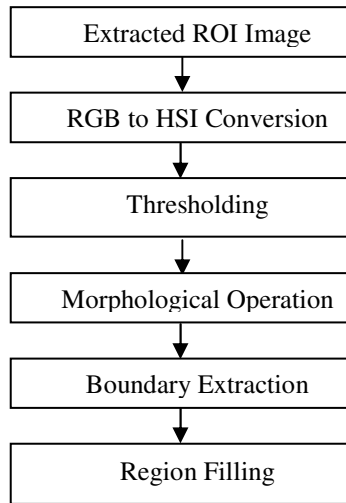


Fig. 2. Block diagram for removing Specular Reflection on ROI cervix images

2.3 Detection of AW Region

There are two stages in the proposed lesion detection method. They are feature extraction stage and classification stage. In the feature extraction stage, the statistical features are extracted from the normal and abnormal region and these features are given to the Bayes classifier for detecting the lesion.

2.3.1 Feature Extraction Stage

The statistical features, mean (μ), standard deviation (σ) and skewness are extracted from each channel of RGB cervix images for every 32x32 overlapping tile. The features of each channel is fused together to form the feature vector. The feature extraction stage is shown in Figure 3. Table 1 shows the average feature values of sample training images

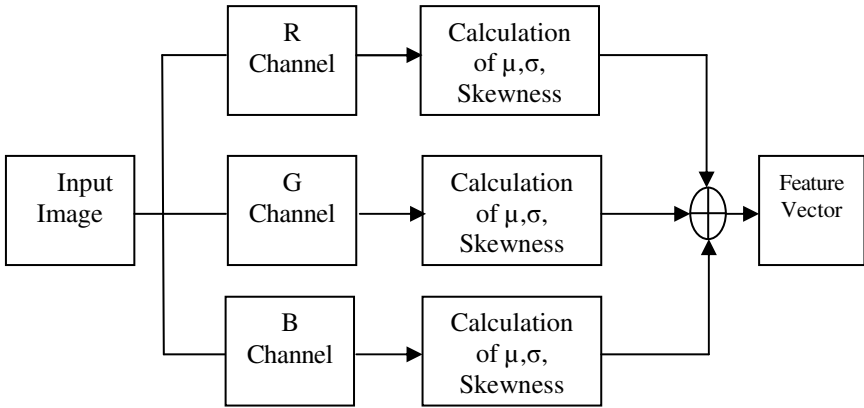


Fig. 3. Feature Extraction Stage

Table 1. Average feature values of sample training images

	Mean		Standard Deviation		Skewness	
	Normal	Abnormal	Normal	Abnormal	Normal	Abnormal
R Channel	93.4302	246.5972	3.5482	0.2509	0.0090	0.3058
G Channel	29.8502	170.2083	1.7766	0.2230	-0.2403	0.1873
B Channel	46.9442	171.3453	2.2051	0.2104	-0.0326	0.2769

2.3.2 Classification Stage

For training the classifier, 25 normal and 25 abnormal cervix images marked by the experts are taken. The normal and abnormal region in cervix images is divided into 32x32 overlapping tiles. The statistical features are extracted from the tiles and stored in the database which will be used for training the classifier. Then the Naïve Bayes classifier is created by fitting the training data in the database. The classification stage is shown in Figure 4.

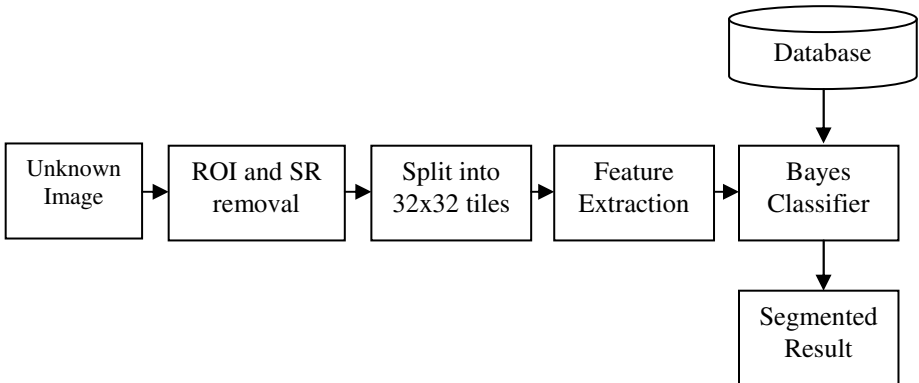


Fig. 4. Classification Stage

3 Experimental Results

The performance of the proposed method is tested on 200 normal colposcopy cervix images and 60 abnormal colposcopy cervix image obtained from Government Kasturibaigandhi Hospital (KGH), Chennai, India. Experimental result of 4 cervix images by the proposed system is shown in Figure 5. The First row image in the figure is normal cervix image. The proposed method correctly classified the normal cervix image as normal and there is no mark by the proposed method and the expert also. The proposed method correctly detects the lesion in the posterior and anterior lib of the second row image and also marked by the experts. The lesions are detected for third row images are similar to the experts' results. The lesions are detected for fourth row images and also similar to the experts' results

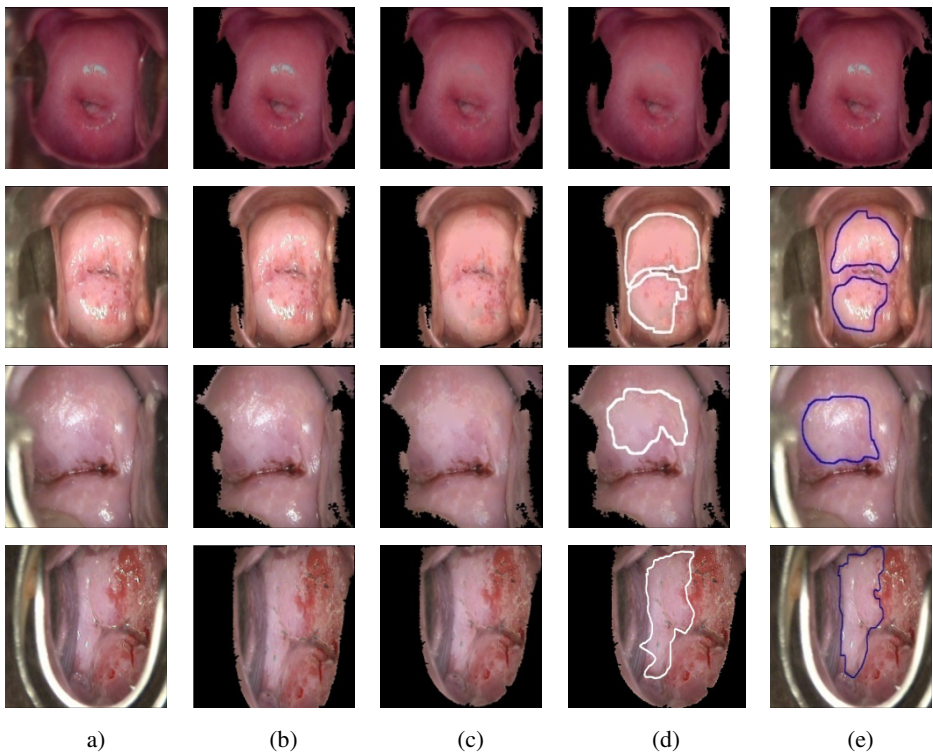


Fig. 5. (a) Input Image (b) ROI segmentation (c) SR removal (d) Proposed method (e) Marked by expert

4 Conclusion

In this paper, a fully automatic method for removal of irrelevant information such as regions outside the cervix, speculum in colposcopy cervix images based on

mathematical morphology and clustering based on Gaussian Mixture Modeling is presented. Also the lesion in the cervix image is detected based on statistical features. The purpose of the recommended process is mainly to prepare the colposcopy cervix images for further analysis. The proposed algorithm is tested on a large set of images totally 240 images and encouraging results are achieved as compared with the experts' results.

In most of the cases, the proposed method detects the lesions very well. However, some times the region outside the lesion edges is also detected as lesion. This is due to the fact that the lesion edges and the region outside lesion edges are classified as abnormal by the classifier based on the features. Further work is in progress to get the better result.

References

1. Claude, I., Pouletaut, P.: Integrated color and texture tools for colposcopic image segmentation. In: IEEE International Conference on Image Processing, pp. 311–314 (2001)
2. Gordon, S., Greenspan, H.: Segmentation of Non-Convex Regions within Uterine Cervix Images. In: IEEE International Symposium on Biomedical Imaging, pp. 312–315 (2007)
3. Greenspan, H., Gordon, S.: Automatic Detection of Anatomical Landmarks in Uterine Cervix Images. IEEE transaction on Medical Imaging, 454–468 (2009)
4. Claude, I., Winzenrieth, R.: Contour features for colposcopic image classification by artificial neural networks. In: IEEE International Conference on Pattern Recognition, pp. 771–774 (2002)
5. Tulpule, B., Yang, S.: Segmentation and Classification of Cervix Lesions by Pattern and Texture Analysis. In: IEEE International Conference on Fuzzy System, pp. 173–176 (2005)
6. Artan, Y., Huang, X.: Combining Multiple 2 ν -Svm Classifiers For Tissue Segmentation. In: IEEE International Symposium on Biomedical Imaging, pp. 488–491 (2003)
7. Acosta-Mesa Héctor, G., Barbara, Z.: Cervical Cancer Detection Using Colposcopic Images: a Temporal Approach. In: IEEE International Conference on Computer Science, pp. 158–164 (2005)
8. Srinivasan, Y., Corona, E.: A Unified Model-Based Image Analysis Framework for Automated Detection of Precancerous Lesions in Digitized Uterine Cervix Images. IEEE Journal Of Selected Topics In Signal Processing, 101–111 (2009)

Heterogeneous Matchmaking Approaches for Semantic Web Service Discovery Using OWL-S

P. Ravinder Reddy¹, A. Damodaram², and A.V. Krishna Prasad³

¹ Department of Computer Science, MIPGS,
JNIAS-JNTUH, Hyderabad, India
message2ravinderi@gmail.com

² Department of Computer Science,
JNTUH, Hyderabad, India
damodarama@jntuh.ac.in

³ Department of Computer Science,
MIPGS, Hyderabad, India
S.V. University, Tirupati, India
kpvambati@gmail.com

Abstract. Services oriented computing is playing a vital role in last decade to develop service oriented distributed computing systems. Web services are reusable software components on the web which can be discovered, fetched, and invoked. With an increase importance towards semantic web services, a challenging task with this domain lies in discovering, composing and then invoking on heterogeneous interface. Matching algorithms are considered basic approach to discover loosely coupled internet registry based web services but algorithms which match based on semantics of the query are limited. Graph based, logic based and many other techniques were introduced to accomplish the task. This paper entitles an overview on different matchmaking approaches for semantic web service discovery using OWL-S.

Keywords: Semantic Web Services, Web Services, SOA, Matching algorithms.

1 Introduction

Distributed systems mediate interactions between clients and applications from heterogeneous platforms. In these interactions, Web service plays vital role. Web service act as a reusable component that can be self described, can be published, and invoked on standard internet protocols. This model of services provide on demand dynamic information to the websites. A Web service is a software system identified by a URI, whose public interfaces and bindings are defined and described using XML. Its definition can be discovered by other software systems [10]. These systems may then interact with the Web service in a manner prescribed by its definition, using XML based messages conveyed by Internet protocols. Several XML standards are been framed like WSDL [8], UDDI [11], SOAP, OWL-S [9], RDF [10].

Wide adoption of web services is due to its simplicity and interoperability it provides over on web. Humans are directly involved in application development

associated with web services. The ultimate vision of SOA is to enable a client to automatically select appropriate service from a pool of dynamically discovered services and invoke it without having the a priori knowledge about service provider and the specifics of the service itself [3]. SOA raises challenging tasks like automatic service discovery, execution time, service composition, service selection, and service invocation.

Semantic web technology realizes machine readable data and domain ontology with the help of open standards like XML, OWL and RDF. Infrastructure of semantic web is adapted to visualize semantic web services.

Matching algorithms are considered basic approach to discover loosely coupled internet registry based web services but algorithms which match based on semantics of the query are limited. In this paper we present an overview on different matchmaking approaches for semantic web service discovery using OWL-S.

The outline of the paper is as follows. After short notes on standards for semantic web services in section 2, we present what actually semantic searching mean, and discuss different approaches for semantic service matching using OWL-S in section3. Comparison on different approaches discussed in section4 and a conclusion on related work in section 5.

1.1 Standards for Semantic Web Services

Ontology are the best suited to add semantics to data on web. An ontology models domain knowledge in terms of concepts and relationships between them [4]. Standard language used to define domain ontology is the web ontology language recommended by World Wide Web consortium. *RDF* data models resources and relations between them [7]. *WSDL* describes the web services in terms of types, Messages, Operation, Port Type, Binding, Port, and Service [8]. *UDDI* is standard registry for publishing, listing out the services over on the web so as to simplify the search done by the consumers based on category, interface, transport and security protocols and specific keyword search. *OWL-S* is the web ontology language for describing web services on owl [9]. OWL-S (formerly DAML-S) is an ontology of services that enables users and software agents able to discover, invoke, compose, and monitor Web resources offering particular services and having particular properties, and should be able to do so with a high degree of automation [6]. OWL-S classifies web services description in to a) *service profiles* b) *service model* c) *service grounding* described below.

a) *Service Profiles* enables service provides to describe services in terms three sections such as provider information, functional description and in the last properties like category of service and quality. Provider information contains contact details. Functional description specifies what service consumes (inputs), service produces (outputs), service conditions should met initially (preconditions) and lastly what it results (effects) after service execution (IOPE).

b) *Service model* can be described as process and helps in interaction for consumers or software agents with the selected web service. Inputs and outputs are subclasses of *parameter*. Preconditions and effects are of kind *expression*, represented as logical formula. Preconditions are to be true in order to execute web service properly. *Atomic process* describes one operation with defined input messages and defined output

result. Group of Atomic process are built to form *Composite process*, that describes sequence of operations (Sequence, Split, Split + Join, Choice, Any-Order, Condition, If-Then-Else, Iterate, Repeat-While, and Repeat-Until) one after the other and maintain some state.

c) *Grounding* describes details regarding how service is accessed pertaining to message formats protocols, serialization, transport and addressing [9]. Concrete level specification description is done in grounding towards the service interaction. While comparing with service profile and service model they exhibit abstract specifications. WSDL is adopted as standard for describing Service Grounding. Since WSDL can deal with network protocols but not with semantic, and OWL-S can deal with semantics but not with network protocols both are overlapped. Mapping OWL-S and WSDL is done

2 Semantic Matchmaking for Web Services

Matching algorithms are search based on keyword matching basically UDDI publishes web services, service provides advertises them and WSDL describes services. Clients or agents (Consumers) usually without prior knowledge of services, search automatically to invoke services. This process may be as simple as similarity search i.e. matching requested parameters with advertising functionalities. But results are ranked based on the degree of match. As per Paolucci [6] matching algorithm should minimize false positives and false negatives, but simple matching increases false positives and false negatives so semantics based matching are introduced to into OWL-S. In semantic matching, meanings are considered with the help of domain ontology. An ontology models domain knowledge in terms of concepts and relationships between them [4]. Functionalities of web services are described in terms of IOPE. Inputs and outputs of service are expressed as concepts belonging to set of ontology [3]. Single Concepts in ontologies refer to many relationships from difference sources. This helps to provide solution for semantic matching of web services. Service provides in describing OWL-S has to be honest in terms of non functional properties like cost and efficient in terms of excess delays.

3 Survey on Matching Algorithms

3.1 Semantic Matching of Web Services Capabilities

Semantic matching is done between the service description being requested and service description being offered by service provider [6] i.e. each concept (both input and output) of request is compared with the concept defined in advertisement using DAML-S (Language for service description). Algorithm initially, outputs of request are matched with outputs of advertisement then similarly inputs of the advertisement is matched with inputs of the request. Resulted retrieved list is given a score based on degree of match (dom) according to the rules mentioned below. Here outR is output requested and outA is output Advertised.

```

degreeofmatch(outR,outA)
    if outA=outR then return exact
    if outR subclassOf outA then return exact
    if outA subsumes outR then return plugIn
    if outR subsumes outA then return subsumes
    otherwise fail.
    
```

3.2 Matching Based on Bipartite Graph

Bipartite graph is special type of graph in which the set of vertices can be divided in to two disjoint subsets, such that each edge connects a vertex from one set to a vertex from other subset. Sample bipartite graph is shown in fig below. Simple Graph $G=(V, E)$ is called bipartite if its vertex set can be partitioned into two disjoint subsets $V=V1 \cup V2$, such that every edge has the form $e=(a, b)$ where $a \in V1$ and $b \in V2$. Note: no vertices both in $V1$ or both in $V2$ are connected.

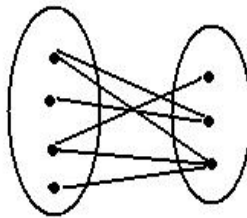


Fig. 1. Bipartite Graph

Same concept of matching is applied in algorithm by [4], where vertex set Q_{out} is Query output and A_{out} is advertisement output. Bipartite graph is constructed $G=(Q_{out} + A_{out}, E)$. Concepts a & b belongs to Q_{out} & A_{out} and R be degree of match (exact, plugin, subsumes, fail) between a and b then edge (a, b) exists if $R \neq fail$ match with weight as degree of match. If output requested is matched with the output advertised then algorithm results in perfect match. When multiple matches are resulted by matching algorithm then for an optimal match Hungarian algorithm technique based on the weights of all edges is used to get minimum value of $max(w_i)$ which is considered as the highest degree of match. Weights of the edges are computed as shown in Table 1. Matching of input concept is also done in similar process, where every advertisement input A_{in} is matched with query input Q_{in} . At the end resulted candidate list is sorted on the basis on input and output concepts queried.

Table 1. Weights calculation table

Degree of match	Weights
Exact	$w1=1$
Plug-in	$w2=(w1 * Q_{out} + 1)$
Subsume	$w3=(w2 * Q_{out} + 1)$

3.3 Flexible Graph Based Approach for Composite Semantic Web Services

In this algorithm [1] for a given composite OWL-S process P, corresponding graph GP is represented. GR be a graph for service requested. The degree of match between two nodes $r \in V(GR)$ and $s \in V(GP)$, respectively are computed for both, inputs and outputs which is as follows

$$dom_{in}(r, s) = \frac{\sum_{u \in IN_s} \max\{sim(u, v)\}}{\|IN_s\|} \quad (1)$$

$$dom_{out}(r, s) = \frac{\sum_{v \in OUT_s} \max\{sim(u, v)\}}{\|OUT_r\|} \quad (2)$$

IN_r and IN_s denote the set of input parameters of r and s , respectively. Function sim computes the similarity between singleton input and output parameters. Similarity function first compares in ontologies defined in OWL-S service description and then string similarity search is done based Cosine Similarity, Jacard Similarity or hybrid measures [5]. Sim function is defined as follows.

$$sim(u, v) = \frac{|\{w \in O \mid u' \sqsubseteq w \wedge u' \sqsubseteq w\}|}{|\{w \in O \mid u' \sqsubseteq w\} \cup \{w \in o \mid v' \sqsubseteq w\}|} \quad (3)$$

Then overall degree of match (4) is computed as follows using Maximum common Subgraph (MCS). The MCS of G1 and G2 is defined as the largest subgraph of G1 that is isomorphic to a subgraph of G2.

$$dom(G_R, G_P) = \frac{\sum_{v \in V_G} dom_v(v, f(v))}{|G_R|} \quad (4)$$

3.4 OWL-MX Hybrid Matching Algorithm

OWL-MX matchmaker [5] is hybrid approach for searching services in OWL-S for given request. Apart from only logic based reasoning this also does content based information retrieval techniques for OWL-S service profile I/O matching with user specification of desire degree and syntactic similarity threshold. For a given service advertisement and request degree of semantic matching is calculated based on the following five filters [5] (first 3 filters are logic other syntactic similarity).

Exact: Service S exactly matches request R $\Leftrightarrow \forall IN_S \exists IN_R : IN_S = IN_R \wedge \forall OUT_R \exists OUT_S : OUT_R = OUT_S$.

Plug-in: Service S plugs into request R $\Leftrightarrow \forall IN_S \exists IN_R : IN_S \supseteq IN_R \wedge \forall OUT_R \exists OUT_S : OUT_S \in LSC(OUT_R)$.

Subsumes: Request R subsumes Service S $\Leftrightarrow \forall IN_S \exists IN_R: IN_S \supseteq IN_R \wedge \forall OUT_R \exists OUT_S: OUT_R \supseteq OUT_S$.

Subsumed-By: Request R is subsumed by Service S $\Leftrightarrow \forall IN_S \exists IN_R: IN_S \supseteq IN_R \wedge \forall OUT_R \exists OUT_S: (OUT_S \supseteq OUT_R \vee OUT_S \in LGC(OUT_R)) \wedge SIM_{IR}(S, R) \geq \alpha$. **Nearest Neighbor:** Service S is nearest neighbour of request R $\Leftrightarrow \forall IN_S \exists IN_R: IN_S \supseteq IN_R \wedge \forall OUT_R \exists OUT_S: OUT_R \supseteq OUT_S \vee SIM_{IR}(S, R) \geq \alpha$.

OWL-MX matching algorithm five variants OWL-M0 to OLW-M4 are implemented. OWL-M0 is logic based service I/O matching and OWL-M1 to OWL-M4 compute the syntactic similarity value of IR for unfolded concepts of query and registered request using loss-of-information measure, extended Jacquard similarity coefficient, the cosine similarity value and the Jensen-Shannon information divergence based similarity value. If the calculated hybrid degree of match is better than or equal to minimum degree specified by user then service is treated as relevant. *Findings:* OWL-MX takes more time in service classifications. Matching of the services depend upon the user specification of degree of match and syntactic similarity threshold.

3.5 Matchmaking Based on Feedback on web Services

In this approach Service feedback are captured from users about the used web Services [2]. Advent of Web2.0 participation and collaboration information helps out to improve accuracy in matchmaking algorithms when searching services. Directly user rating or inference rating of service with user behavior feedback loop component is incorporated in to architecture for capturing used services feedback into RDF [2].

```
<r: Rating>
<foaf: Person rdf: about="#mark"/>
<r: Request rdf: about="#requestX"/>
<r: Service rdf: about="#serviceY"/>
<r: Score rdf: datatype="&xsd; double">0.90</r: score>
</r:Rating>
```

For a given request R a tuple set $T=\{U,R,S,f\}$ where U denotes user rated for request R and service S with scoring f is constructed from rating database $\tau \subseteq U \times R \times S \times F$. feedback score vector fb is calculated as

$$fb(R, S) = \frac{\sum_{(U, Q, S, f) \in \tau: Q \in SIM(R)} f * sim(R, Q)}{|\{(U, Q, S, f) \in \tau: Q \in SIM(R)\}|} \tag{5}$$

Where $sim(R, Q)$ is match instance of Q with respect to R. OWL-MX [7] matching algorithms discussed in above section are used to calculate the matching with all M0, M1, M2, M3 and M4 filters. Match instance for service S, request R, and matching function is vector such that

$$S[i, j] = \begin{cases} \max\{mi(S, p_k, R, p_j), \forall j: p_j \in R_{IN}\} \\ p_k \in S_{IN} \\ \max\{mi(S, p_k, R, p_j), \forall j: p_j \in R_{out}\} \\ p_k \in S_{OUT} \end{cases} \quad (6)$$

Let $u, v \in \tau$ being match instance set of resulted services. u dominates v , $u > v$, iff u has higher degree of match in all parameters for a given request. Dominance scores are calculated to rank matching services based on higher degree of match.

Both feedback score fb as an additional ranked match instance and fb as integrated with ranked match instances strategies are evaluated for matching result.

4 Comparison of All Approaches

In this section we compare present different discussed approaches above

Table 2. Result Comparison table for approaches

	Match Making	Test Collection	Result
OWL-MX	OWL-MX	OWL TC v2	Logic + all hybrid OWL-MX match makers are outrun by IR based service retrieval in terms of average query response time
Feedback	OWLS-MX	OWL TC v2	Feedback Aware methods outperform the non feed back methods
Bipartite	Java, Protege,	OWL TC v1	In terms of performance bipartite matching algorithm is consuming more search time with respect to number of advertisements in registry.
Greedy	DAML-S DAML+OIL Reasoner	AdvDB, OntologyDB	dom are dependent on the concept defined in the profiles. If the order is changed the results of the algorithm results a change in candidate list. Sometimes false positives and false negatives are generated

Table 3. Comparison table for approaches

	I/O	User Pref	Approach	Ontology Type
OWL-MX	Yes	Yes	Logic and content based	Heterogeneous
Feedback	Yes	Yes	Logic and User feedback	Heterogeneous
Flexible	Yes	No	Graph and IR Mechanism	Domain Specific
Bipartite	Yes	No	Bipartite Graph Matching & Hungarian Algorithm	Domain Specific
Greedy	Yes	No	Greedy	Domain specific

5 Conclusion

We have seen different approaches in semantic matchmaking for service discovery. All the algorithms mainly concentrate on function parameters to match. Non functional parameters (price, negotiations, trust, quality) the web2.0 user participation, web service rating system, feedback from user can also be used to narrow down service discovery.

References

1. Cuzzocrea, A., Fisichella, M.: A Flexible Graph-based Approach for Matching Composite Semantic Web Services. In: LWDM 2011, pp. 1–2 (2011)
2. Averbakh, A., Krause, D., Skoutas, D.: Exploiting User Feedback to Improve Semantic Web Service Discovery. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 33–48. Springer, Heidelberg (2009)
3. Bellur, U., Vadodaria, H., Gupta, A.: Semantic Match making Algorithms Chapter 26 in Advances in Greedy Algorithms book, pp. 481–502 (2008)
4. Bellur, U., Kulkarni, R., Rekhi, K.: Improved matchmaking algorithm for semantic web services based on bipartite Matching. In: IEEE International Conference on Web services. ICWS 2007, pp. 86–93 (2007)
5. Klusch, M., Fries, B., Sycara, K.P.: Automated semantic web service discovery with OWLS-MX. In: AAMAS, pp. 915–922 (2006)
6. Paolucci, M., Kawamura, T., Payne, T.R., Sycara, K.: Semantic Matching of Web Services Capabilities. In: Horrocks, I., Hendler, J. (eds.) ISWC 2002. LNCS, vol. 2342, pp. 333–347. Springer, Heidelberg (2002)
7. W3c recommendations OWL Web Ontology Language Overview (2004), <http://www.w3.org/TR/2004/REC-owl-features-20040210/>
8. W3c recommendations Web Services Description Language (WSDL), <http://www.w3.org/TR/wsd1>
9. W3c recommendations OWL-S: Semantic Markup for Web Services, <http://www.w3.org/Submission/OWL-S/>
10. W3c recommendations Web Services Architecture Requirements (2004), <http://www.w3.org/TR/wsa-reqs/>
11. Clement, L., Hately, A., von Riegen, C., Rogers, T.: (2004), <http://www.UDDI/Version3.0.2>
12. RDF Working Group: Resource Description Framework (RDF) (2004), <http://www.w3.org/2001/sw/wiki/RDF>

Design and Implementation of Affective E-Learning Strategy Based on Facial Emotion Recognition

Arindam Ray¹ and Amlan Chakrabarti²

¹ Awadh Centre of Education
Guru Gobind Singh Indraprastha University,
New Delhi, India

arindamray_2007@yahoo.co.in

² A K Choudhury School of Information Technology
University of Calcutta
Kolkata, West Bengal, India
acakcs@caluniv.ac.in

Abstract. E Learning is emerging as a heavily learner-centric, emphasizing pervasive and personalized learning technology. Affective learning outcomes in a nutshell, involve attitudes, motivation, and values. In the same tune we can also define the affective E-learning, as a strategy, which implies recognition of learner's emotion and selection of pedagogy in a best possible way. For the best delivery, learner's affective state needs to be identified where the key solution is emotion recognition. Our work focuses on emotion detection using biophysical signals which further explores the evolution of emotion during learning process, to generate a feedback that can be used to improve learning experiences. Our research is deeply focused into the aspects of operative content delivery mechanism by using physiological facial signals for the detection of learner's emotion but without detecting the face. In this paper we propose a key technique to detect learner's facial expression, based on neural network classification and selection of appropriate learning style, which shows reasonable results in comparison with the other existing systems. The result manifests that the recognizer system is effective.

1 Introduction

A fundamental tenet of this design is that one method does not fit to all learners. Different pedagogy has to be chosen for different learner. In E-Learning portal the method of teaching-learning is unidirectional which implies simultaneous communication can't happen. But in the face to face interactive session, it happens. Teacher's experience plays an important role and hence an E-Learning portal needs such platform for emotion sharing between the learner and the teacher. Learner's emotion first reflects on the face and hence facial emotion recognition [1] is preferred to get the affective state of learner. The proposed model can recognize learners' emotion to identify the affective state. In this paper we propose a technique to detect learner's facial expression using SVM (Support Vector Machine) and also selection of the course based on neural network. As per the psychological theory that human emotions –could be classified into six typical emotions [2] viz. ‘‘happiness’’,

“sadness”, “surprise”, “fear”, “disgust” and “anger”. For the appropriate learning pedagogy, we need to identify the learner’s psychology in the best way. Number of parameters is involved for psychological emotion, but our work is to utilize the facial gesture which can be correlated with emotions using neuro-fuzzy approach of classification. After identifying the psychological emotion, our system will automatically detect the learning pedagogy. The required course detection will be held automatically as per the algorithm. Our total proposal fastens in twofold operations; one is identification of learners’ emotion and another is selection of the learning style.

James J. Lien [3] has shown the process of automated facial expression of upper face. But the detection technique used in the Facial Action Coding System (FACS) [4] to identify facial action was an anatomically based coding system that enables discrimination between closely related expressions. FACS divided the face into upper and lower face action and further subdivided motion into action units (AUs). Their approach recognized upper face expressions in the forehead and brow regions only. For emotion detection lower face also plays an important role and so this approach fulfills partial emotion extraction. Moreover it also needs high gradient image but in web cam, assuming that the learner having a cheap and minimum amount of computational resource, we can’t get such quality images. C.H. Messom, A Sarrafzadeh, M.J. Johnson, F. Chao said in their paper [5] that many software systems would significantly improve performance if they could adapt to the emotional state of the user, for example if intelligent tutoring systems, information / help kiosks, ATM’s, automatic ticketing machines could recognize when users were confused, frustrated or angry then they could guide the user back to remedial help systems, so improving the services. Current software systems are not able to estimate the affective state of the users and so, that are not able to offer these additional capabilities. They have proposed a model for detection of affective state estimator. This system consists of two neural network classifiers and a fuzzy logic facial analysis system. This system has been successfully prototyped for use in an intelligent tutoring system that has been adapted to the affective state of the user. Though facial expressions are the most important means of detecting emotions, however, other bio-signals such as heartbeat, skin resistance and voice tone can also be used for detecting human emotions. Our proposed model detects the movement of some points / spots of the face, from which it can detect the learner’s emotion. Moreover our approach is one step forward i.e. it could select the appropriate course for the learner using neural network.

Our research claims better facial emotion detection of a learner in an E-Learning platform and it requires a very minimum amount of resource at client side for detection of the affective state of the learner. Our objective is to detect affective state of learner through facial emotion; hence we choose the spot detection technique which has not been done earlier. The achievements of our research can be briefed as follows:

- It works in client side, which minimizes the server side overhead.
- It doesn’t needs to detect the full face; only spot movement detection is enough for emotions detection.
- Emotion can easily be detected from the same group of learners because of the similar facial pattern.
- There is a model for automatic lesson detection.

Our proposed work accomplishes a fusion of facial emotion and learning pedagogy, ensuring an affective E-Learning strategy, which is a new work in this domain of research.

In this paper, the abstraction of the affective computing model in E-Learning that identifies the effectiveness of learner has been explained in Section 1.1. Learner’s facial expression capturing and emotion detection framework has been implemented in the section 2. The key technology and methodologies for facial emotion recognition has been implemented in Section 3. Implementation and results of the facial emotion recognition using SVM has been shown in Section 4 and conclusive remarks are given in Section 5.

1.1 E-Learning Model Based on Affective Computing

Mase K. proposed an emotion recognition system that has used the major directions of specific facial muscles [6]. We are proposing the model for affective computing, based on fusion of emotional behaviors (speech and facial expressions) which works as an important feedback signal to know about psychological state of the learner. Feedback signals can be taken care effectively and can help in tuning the teaching strategies to serve personalized learning. We have taken the traditional E-Learning model and have added the affective computing module with it. The proposed model of E-learning system based on affective computing is shown in Figure 1.

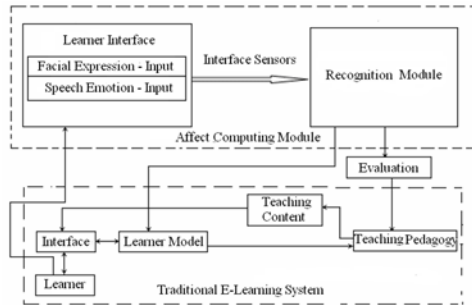


Fig. 1. E-Learning Model Based on Affective Computing

Learner Interface module: Affective computing input (speech emotion and facial expression recognition input) is given to the human machine interface of the traditional E-Learning system, which collects learners’ speech and facial emotion feedback information primarily, and thus realizes the emotion compensation. In this paper we have focused on facial expression only.

Recognition module: Emotion recognition module is composed of input, pre-processing, feature extraction, feature selection and emotion recognition sub modules.

Evaluation module: It takes the input from the recognition module and generates the corresponding evaluation parameters.

2 Facial Emotion Recognition By Capturing Learners' Facial Expression

2.1 Technical Framework

The recognition process has been framed with intelligent affective state recognizer. This affective structure is being composed with scanning of Images, Pre-Processing, Classification, Feature Extraction and Interpretation [7]. The following steps have been executed for the task:

Step I: Image of the learner is captured before the course delivery and stored in the database.

Step II: The image of the learner is again captured after the course delivery to get the changes in the face emotion.

Step III: The image acquired in Step II is then pre-processed by a neural network so that the positions of the face, specially forehead, eyebrow, low eye, cheek areas are detected.

Step IV: This phase identifies the user, based on the database of known user image using Support Vector Machine.

Step V: This stage consists of a facial feature extraction system based on neural network and an affective state estimator based on fuzzy logic.

Step VI: This final phase selects the course delivery module as per the learners' learning style.

The extraction of facial features could be done by the use of markers, so we don't use face detection and tracking algorithms. All of the processing will be completed at the client side, which means that the online system will only send the final affective state of the user to the server side of the system which requires only a small overhead.

3 Key Technology Adopted for Emotion Detection

3.1 Training Data

Our training dataset consists of 129 college students and staffs ranging in age between 19 to 35 years. 45% were female, 55% were male and all were Indians. Videos were recorded in QHM495LM-3207 web camera located directly in front of the learner whose basic configuration are - lens capacity:14 Mega pixels; output Size: 640x480; capture size 640x480; avoid flicker 50Hz; zoom 1x. Subjects have been delivered by an experimenter to perform a series of all 129 facial expressions. Image sequences from neutral to target display were 640 by 480 pixel arrays with 8-bit precision. The only selection criterion was that a sequence be labeled as one of the 6 basic emotions (disgust, sadness, happiness, fear, anger, surprise). The sequences came from 5 different subjects, with 1 to 6 emotions per subject and we have classified the emotions in ten groups for further course selection.

3.2 Methodology Adapted

Facial expressions give important clues about emotions. Therefore, the features used are typically based on local spatial position or displacement of specific points and regions of the face. For a complete review of recent emotion recognition systems based on facial expression the readers are referred to [8]. A motion capture system (Webcam) was used to capture the expressive facial motion after the delivery of the course. Notice that the facial features are extracted with the precision of webcam (14 Mega pixels).

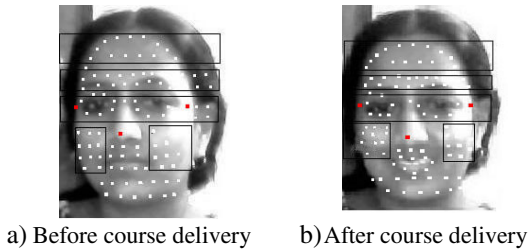


Fig. 2. Five areas of the face has been considered in this study

In the system based on visual information, the spatial data collected from markers in each frame of the video were reduced into a 6-dimensional feature vector per sentence, which is used as input to the classifier. The facial expression system, which is shown in Figure 2, is described below. After capturing the motion data, it has been normalized.

- (1) All markers are translated in order to make a nose marker which acts as the local coordinate center of each frame,
- (2) One frame with neutral and close-mouth head pose is picked as the reference frame,
- (3) Three approximately rigid markers (manually chosen and illustrated as red points in Figure 2) define a local coordinate origin for each frame, and
- (4) Each frame is rotated to align it with the reference frame. Each data frame is divided into five blocks: *forehead*, *eyebrow*, *low eye*, *right cheek* and *left cheek area*.

For each block, the 3D coordinate of markers in this block is concatenated together to form a data vector. It has been noticed that the markers near the lips are not considered, because the articulation of the speech might be recognized as a smile, which will confuse the emotion recognition system [9]. It is well observed that the different emotions appear in separate clusters, so important clues could be extracted from the spatial position of these 6-dimensional features space. Psychological research has classified six facial expressions which correspond to distinct universal emotions [10]. It is interesting to note that four out of the six are negative emotions. We have generalized the cues for facial expression as given below [11].

Table 1. Facial Expression and its motion cues

Expression	Motion Cues	Expression	Motion Cues
Happiness	raising and lowering of mouth corners	Fear	brows raised eyes open mouth opens slightly
Sadness	lowering of mouth corners raise inner portion of brows	Disgust	upper lip is raised nose bridge is wrinkled cheeks raised
Surprise	brows arch eyes open wide to expose more white jaw drops slightly	Anger	brows lowered lips pressed firmly eyes bulging

4 Implementation and Results

We have used the support vector machine (SVM) [12] concept for setting of related supervised learning methods that analyze data and recognized patterns, used for classification and regression analysis. The facial spots in the individual five blocks have been identified before and after the course delivery. Our objective is to classify data and hence our training data points belong to one of the two classes and the goal is to decide which class a new data point will be. SVM views as a p -dimensional vector (a list of p numbers), and we found separated points in a $(p - 1)$ dimensional hyperplane. In our training data we have used 81- dimensions, so $p=81$. One reasonable choice as the best hyperplane is the one that represents the largest separation, or margin, between the two classes. So we have chosen the hyperplane so that the distance from it to the nearest data point on each side is maximized. This resulted to the hyperplane, the perceptron (ANNs) of optimal stability.

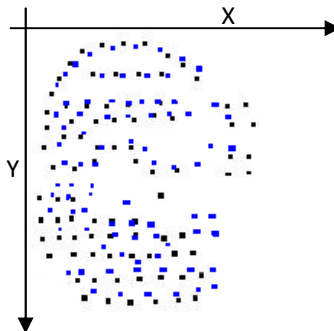


Fig. 3. Facial points before and after course delivery

We have extracted the set of values from the facial spots from one training sample. We have identified 81 spots throughout the face and as per the muscle movements and it has been divided into the 5 blocks. Before the delivery of the course snap was taken (Figure 2(a)) and was marked in black spot set. After the delivery of the course another snap was taken (Figure 2(b)) and those marks were done with blue spot set.

The both are mapped for detection of pattern using the SVM as shown in Figure 3. Now we go ahead for assigning the prediction class. The black points are considered in one class (+1) and the blue points are considered in another class (-1). We have taken a training data, which is a set of n points.

Positive: $\langle \mathbf{w} \bullet \mathbf{x} \rangle + b = +1$ Negative: $\langle \mathbf{w} \bullet \mathbf{x} \rangle + b = -1$ Hyperplane: $\langle \mathbf{w} \bullet \mathbf{x} \rangle + b = 0$

We have to find the unknowns, \mathbf{w} and \mathbf{b} by expending the equations:

$$w_1x_1 + w_2x_2 + b = +1 \tag{1}$$

$$w_1x_1 + w_2x_2 + b = -1 \tag{2}$$

$$w_1x_1 + w_2x_2 + b = 0 \tag{3}$$

We have generated the training data for 129 facial emotions before and after course delivery. One dataset content 81 values of which first ten rows of the training data is given in Table 2. For all 129 training snaps same set of tables has been generated.

Table 2. Training data (First 10 rows)

Sample	Black (+1)			Blue (-1)			Sample	Black (+1)			Blue (-1)		
SL	Class	X	Y	Class	X	Y	SL	Class	X	Y	Class	X	Y
1	+1	15	71	-1	17	65	6	+1	61	37	-1	68	39
2	+1	20	61	-1	23	58	7	+1	75	38	-1	85	41
3	+1	25	51	-1	30	57	8	+1	91	46	-1	100	47
4	+1	34	42	-1	39	44	9	+1	100	52	-1	111	54
5	+1	47	38	-1	54	41	10	+1	108	61	-1	47	58

By using DTREG-SVM (www.dtreg.com) modeling we have generated the report which is shown in Table 3

4.1 Analysis of the Report

From one training data, the result is presented in Table 3.

Table 3. Result of the training data

Bin Index	Cutoff Target	Mean Predicted	Mean Actual	Cum % Population	Cum % Target	Cum % Gain	% of Population	% of Target	% of Lift
1	1.9971709	1.9972103	2.0000000	11.11	14.81	1.33	11.11	14.81	1.33
2	1.9971441	1.9971503	2.0000000	22.22	29.63	1.33	11.11	14.81	1.33
3	1.9971144	1.9971161	2.0000000	33.33	44.44	1.33	11.11	14.81	1.33
4	1.9971038	1.9971082	2.0000000	44.44	59.26	1.33	11.11	14.81	1.33
5	1.0030278	1.4999917	1.5000000	55.56	70.37	1.27	11.11	11.11	1.00
6	1.0030010	1.0030055	1.0000000	66.67	77.78	1.17	11.11	7.41	0.67
7	1.0029607	1.0029694	1.0000000	77.78	85.19	1.10	11.11	7.41	0.67
8	1.0028067	1.0028810	1.0000000	88.89	92.59	1.04	11.11	7.41	0.67
9	1.0023318	1.0025693	1.0000000	100.00	100.00	1.00	11.11	7.41	0.67
10	1.0023318	0.0000000	0.0000000	100.00	100.00	1.00	0.00	0.00	0.00
Average gain = 1.190				Mean value of target variable = 1.5					

4.2 Explanation

We need the average gain for further identification of the course delivery pattern. The rest of the data is not required in our study. As per the software, if the gain is 1.00 or less implies doesn't detect the pattern and hence we have taken the average of all the training samples. First 10 training sample average gains are shown in Table 4.

Table 4. First 10 Training Data

Sample No	Average Gain	Sample No	Average Gain
1	1.1900	6	1.1200
2	1.9100	7	1.1800
3	1.0000	8	1.1700
4	1.2900	9	1.1403
5	1.1100	10	1.4404

Average of all 129 training sample is: *1.3693* and which is more than 1.00, this implies that the pattern has been recognized. The next step is to make a group for lesson identification and as per the psychological theory the groups are formed which is given in Table 5.

Table 5. Emotion - Group – Decision Taken

Emotion Detected	Grouping	Learning Style detection
Happiness, surprise	Positive Group	Lesson Understood
Sadness, fear, Disgust, anger	Negative Group	Lesson Not Understood

4.3 Modeling of Learning Styles with Neural Networks

E-learning environments can take advantage of these different forms of learning by recognizing the pedagogy of each individual student using the system and adapting the content of courses to match this style. The method is based on artificial neural networks (ANNs) [13]. Neural networks are computational models for classification inspired by the neural structure of the brain: models that have proven to produce very accurate classifiers. In the proposed approach, neural networks are used to recognize learners' learning styles based upon the actions they have performed in an E-Learning system. As per the detection of affective state, the system will suggest the lesson to the individual learner. As per the flow diagram system can select the learning pedagogy of the learner and neuro-fuzzy logic is required to implement such methodology. The flow diagram is shown in Figure 4

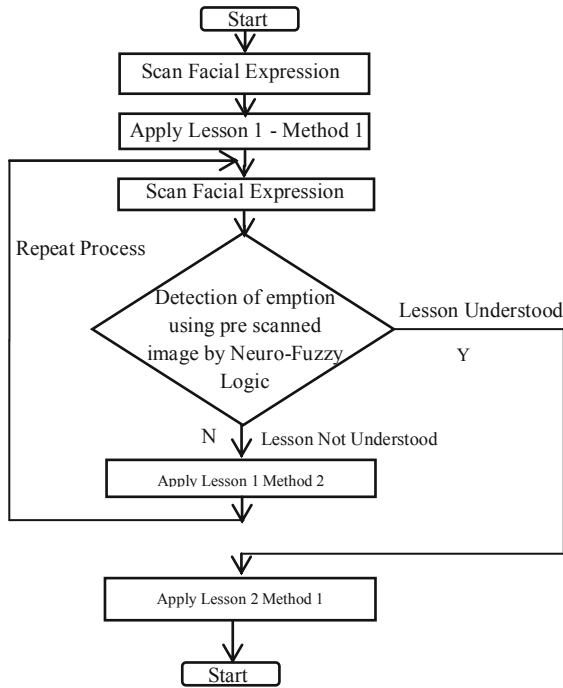


Fig. 4. Flow of Learning Method identification

5 Conclusion

In this paper, we have shown the effectiveness of facial emotion recognition in order to identify the affective state of a learner. This research also analyzed the strengths of facial expression classifiers in E-Learning environment using SVM. The results presented in this research shows that it is feasible to recognize human affective states with high accuracy by the use of visual modalities. Therefore, the next generation of human-computer interfaces might be able to perceive humans feedback, and respond appropriately and opportunely to get users' affective states, improving the performance and engagement of the current interfaces.

In future work, the other image capturing obstructions like lateral impression, face with spectacle, sweaty face, etc. are to be considered. Our goal is to increase the maximum competence in human computer interaction for building up affective E-Learning system.

References

1. Cacioppo, J.T., Tassinary, L.G.: Inferring psychological significance from physiological signals. *American Psychologist*, 16–28 (1990)
2. Elfenbein, H.A., Ambady, N.: Universals and cultural differences in understanding emotions. *Curr. Dir. Psychol. Sci.* 12(5), 159–164 (2003a)

3. Lien, J.J., Kanade, T., Cohn, J.F., Li, C.-C.: Automated Facial Expression Recognition Based on FACS Action Units. In: IEEE 3rd International Conference, pp. 390–395 (1998)
4. Ekman, P., Friesen, W.V.: The Facial Action Coding System. Consulting Psychologists Press, Inc., San Francisco (1978)
5. Messom, C.H., Sarrafzadeh, A., Johnson, M.J., Chao, F.: Affective State Estimation From Facial Images Using Neural Networks And Fuzzy Logic, http://www.massey.ac.nz/~chmessom/Manuscript%20NGITS_NN.pdf
6. Mase, K.: Recognition of facial expression from optical flow. IEICE Trans., E. 74(10), 3474–3483 (1991)
7. Zhu, E., Liu, Q., Xu, X., Lei, T.: Research on Affective State Recognition in E-Learning System by Using Neural Network. In: Shi, Y., van Albada, G.D., Dongarra, J., Sloat, P.M.A. (eds.) ICCS 2007, Part III, LNCS, vol. 4489, pp. 575–578. Springer, Heidelberg (2007)
8. Pantic, M., Rothkrantz, L.J.M.: Toward an affect-sensitive multimodal human-computer interaction. Proceedings of the IEEE 91(9), 1370–1390 (2003)
9. Guo, G., Li, S., Chan, K.: Face recognition by support vector machines. Image and Vision Computing 19(9-10), 631–638 (2001)
10. Yacoob, Y., Davis, L.: Recognizing Human Facial Expressions from Long Image Sequences Using Optical Flow. IEEE Transactions on Pattern Analysis and Machine Intelligence 18(6), 636–642 (1996)
11. Ekman, Friesen: Universals and cultural differences in the judgments of facial expressions of emotion. Journal of Personality and Social Psychology 53(4), 712–717 (1987)
12. Burges, C.: A tutorial on support vector machines for pattern recognition. Data Mining and Know. Disc. 2(2), 1–47 (1998)
13. Keefe, L.W.: Learning style: an overview. NASSP's Student learning styles: Diagnosing and prescribing programs, 1–17 (1979)

A Relaxed Parzen Window Based Multifeatured Fuzzy-GIS Model to Forecast Facility Locations (RPWMFGISFFL)

Parthajit Roy¹ and Jyotsna Kumar Mandal²

¹ Department of Computer Science,
The University of Burdwan, Burdwan, West Bengal,
India-713104

roy.parthajit@gmail.com

² Department of Computer Science and Engineering,
Kalyani University, Kalyani, West Bengal, India-741235
jkm.cse@gmail.com

Abstract. In this paper, a Relaxed Parzen Window based Multifeatured Fuzzy-GIS model has been proposed to forecast future facility locations for various community services. Delaunay Triangulation technique is used as the base of the model. Given a set of facility locations in a locality, the model forecasts the best feasible places for the introduction of new facility centers based on various attribute values. The Decision Support System of the proposed model is based on Relaxed Parzen Window and Fuzzy implication technique. The degree of need depends on the distance vector and the expected population of the locality.

Keywords: RPWMFGISFFL, Delaunay Triangulation, Parzen Window, Fuzzy-GIS, Forecasting of Facility Locations.

1 Introduction

Geographical Information System [GIS], has become very trendy now a days[8]. In this paper, a Relaxed Parzen Window based multi-featured Fuzzy-GIS model has been proposed for forecasting the places where the new facility locations (Health Center in present case) can be established. The model considers Delaunay Triangulation(DT) method [1] for first level computations. Some applications of Delaunay Triangulation in GIS is available in various literatures[7][10] and some good online resources are made available by L. Paul Chew [3] and Chris Gold [6] which deals with Computational Geometry based problems.

In the second level, a Parzen Window(PW) [4] based Fuzzy conjunction concept [11] for the prediction of degree of need has been proposed. Instead of crisp Parzen Window, the model used a Relaxed one to handle the need realistically. The proposed model is based on a need based fuzzy implication model. Some application of Fuzzy Logic in GIS has been discussed by Gavin Fleming et.al.[5].

Section 2 of the paper deals with the outline of the proposed model. The detail of anatomy of the model has been illustrated in Section 3. Experimental result and discussions are given in Section 4. Conclusions are given in Section 5 and references are drawn at the end.

2 Scheme

The proposed scheme has three phases. These are as follows:

- Find places for new facility centers using Delaunay Triangulation [DT] considering the Near-neighbor distance vector.
- Project population of the suggested location using Fuzzy Logic and DT.
- Suggest the need of the new facility center based on Relaxed Fuzzy Parzen Window model.

The Phases are discussed below:

In the first phase, a Delaunay Triangulation based Facility Center Projection (DT-FCP) model has been proposed, which decomposes the *map of interest* into a set of triangles based on existing facility locations. Utilizing these triangles the model proposes the locations, where the facility centers can be established.

The second phase computes Delaunay Triangulation based Projection for Population (DTPP) of the suggestion points considering available population information.

Taking the expected population of the suggestion point, the model applies a Parzen Window based Fuzzy model to suggest the need of the facility center at that point. The degree of need is calculated by relaxing the window size of the PW.

Algorithm 1, Algorithm 2 and Algorithm 3 are techniques pertaining to the proposed system and outlines the model. The anatomy of the model is described in section 3 by exploring the underlying geometrical and fuzzy computation model.

Algorithm 1. Location Selection

Input: $S = \{f_1, f_2, \dots, f_n\}$ ▷ Existing facility centers within boundary map
Input: $B = \{l_1, l_2, \dots, l_m\}$ ▷ Boundary lines of the Map.
 1: **place** $S = \{f_1, f_2, \dots, f_n\}$ on the Map.
 2: **compute** Delaunay Triangles of Facility Centers [DTFCP].
 3: **compute** $T \leftarrow$ Triangle with Largest circumcircle
 4: **compute** $c \leftarrow$ circumcenter of T
 5: **if** $c \in$ Inside Boundary **then**
 6: **return** c , radius ▷ Returns the location with maximum distance from nearest facilities
 7: **end if**

Algorithm 2. Population Prediction

Precondition: $S = \{(p_1, k_1), (p_2, k_2), \dots, (p_t, k_t)\}$ ▷ t different points with known population.
Precondition: $B = \{l_1, l_2, \dots, l_m\}$ ▷ Boundary lines
Input: $c(x, y)$ ▷ Point suggested by the Algorithm 1
 1: **compute** Delaunay Triangles of Population Points [DTPP] ▷ DTPP
 2: **find** the triangle T s.t. $c(x, y) \in T$ ▷ Triangle containing point suggested by Algorithm 1
 3: **compute** the population of $c(x, y)$ based on fuzzy distance vector. ▷ Considering corners of T
 4: **return** population of $c(x, y)$

Algorithm 3. Relaxed Parzen Window based Fuzzy Need Calculation

- 1: **call** Location Selection algorithm. Algo 1 ▷ We shall get the point $c(x,y)$ with largest distance from nearest facility locations.
- 2: **call** population prediction algorithm. Algo 2 ▷ Calculate the population of the point $c(x,y)$
- 3: **pass** these two to relaxed parzen window model and get the need of facility center at that point
- 4: **suggest** the need in the visual form on the map with different legends.
- 5: **add** the point $c(x,y)$ in the existing Delaunay Triangulation[DTFCP] and recompute it.
- 6: **repeat** the same procedure so that all possible locations are explored one after another.

3 The Model of Computation

To find the solution, two computational models have been adopted. Delaunay Triangulation [DT] and Parzen Window [PW] based Fuzzy implication. A DT is a planner decomposition of a set of points where the edges forms triangle. In the present paper Bowyer-Watson algorithm has been considered for computing DT [2] [9].

Firstly the model computes the DT of the set of existing facility centers. Then it calculates the radius of all of the circum circles of the set of triangles [Fig 1(a)]. The triangle with largest radius is considered first. As the circumcircle of a Delaunay Triangle is empty, the circum center of the largest triangle is the point which is furthest from all other facility locations and is the most deprived location in the map. If this point lies inside the boundary, then it is the best location for a new facility center.

The model then checks the population of the suggested location using fuzzy k -nearest neighborhood method in the following way:

For every small locations, the population is represented by a point within the location, called the representative point. The model then considers the Delaunay Triangulation of the existing population points[Fig 1(b)]. Clearly, the suggestion point will fall inside some of the DT of the population DT set. The model tracks that triangle and computes population of the point using fuzzy k -nearest neighborhood algorithm for $k=3$. The computation is done in the following way:

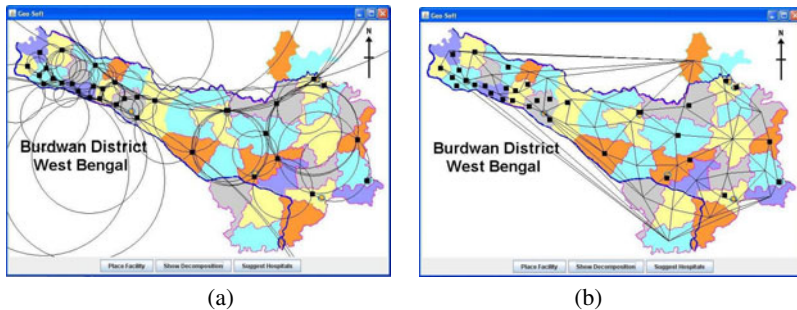


Fig. 1. (a) The circumcircles of DT of existing facility locations. (b) The DT of known population points.

Let three corners of a population Delaunay triangle are $\mathcal{A}(x_1, y_1)$, $\mathcal{B}(x_2, y_2)$ and $\mathcal{C}(x_3, y_3)$ [Fig 2(a)]. Also let the point inside the triangle i.e. the suggested point is $\mathcal{P}(x_k, y_k)$. Let $\mathcal{D}(\xi, \zeta)$ is the intersection point of extended $\overline{\mathcal{A}\mathcal{P}}$ and $\overline{\mathcal{B}\mathcal{C}}$. Let the populations at \mathcal{A} , \mathcal{B} and \mathcal{C} are known and are $\Psi_{\mathcal{A}}$, $\Psi_{\mathcal{B}}$ and $\Psi_{\mathcal{C}}$ respectively. Clearly, the point $\mathcal{P}(x_k, y_k)$ can be represented as a convex combination of (as it lies inside the triangle) $\mathcal{A}(x_1, y_1)$, $\mathcal{B}(x_2, y_2)$ and $\mathcal{C}(x_3, y_3)$. For this, we first find the location of the point \mathcal{D} . Using coordinate geometry $\mathcal{D}(\xi, \zeta)$ can be deduced as:

$$\xi = ((\frac{y_k - y_1}{x_k - x_1})x_1 - (\frac{y_3 - y_2}{x_3 - x_2})x_2 + y_2 - y_1) / (\frac{y_k - y_1}{x_k - x_1} - \frac{y_3 - y_2}{x_3 - x_2}) \text{ and}$$

$$\zeta = ((\frac{x_k - x_1}{y_k - y_1})y_1 - (\frac{x_3 - x_2}{y_3 - y_2})y_2 + x_2 - x_1) / (\frac{x_k - x_1}{y_k - y_1} - \frac{x_3 - x_2}{y_3 - y_2})$$

Let α is the ratio $\frac{\overline{\mathcal{B}\mathcal{D}}}{\overline{\mathcal{B}\mathcal{C}}}$ and β is the ratio $\frac{\overline{\mathcal{A}\mathcal{D}}}{\overline{\mathcal{A}\mathcal{C}}}$. Now, Euclidian distance formula gives,

$$\alpha = \frac{\sqrt{(x_2 - \xi)^2 + (y_2 - \zeta)^2}}{\sqrt{(x_2 - x_3)^2 + (y_2 - y_3)^2}} \text{ and } \beta = \frac{\sqrt{(x_1 - x_k)^2 + (y_1 - y_k)^2}}{\sqrt{(x_1 - \xi)^2 + (y_1 - \zeta)^2}}$$

So, population of the point $\mathcal{D}(\xi, \eta)$ (denoted by $\Psi_{\mathcal{D}}$) on the line $\overline{\mathcal{B}\mathcal{C}}$ can be given by the linear combination of population of \mathcal{B} and \mathcal{C} as follows:

$\Psi_{\mathcal{D}} = \alpha\Psi_{\mathcal{C}} + (1 - \alpha)\Psi_{\mathcal{B}}$. Similarly for the point \mathcal{P} on the line $\overline{\mathcal{A}\mathcal{C}}$, the linear combination is, $\beta\Psi_{\mathcal{C}} + (1 - \beta)\Psi_{\mathcal{A}}$ i.e. $\beta(\alpha\Psi_{\mathcal{C}} + (1 - \alpha)\Psi_{\mathcal{B}}) + (1 - \beta)\Psi_{\mathcal{A}}$

The above mentioned technique suggests a fuzzy method for projecting the population of the suggestion point. The system opens a Parzen Window[PW] to decide the degree of need of that point.

Parzen Window[PW] is an approach of pattern classification where a d dimensional hypercube is considered [4]. The length of an edge of hypercube is h_n and d is the number of features considered. The Volume of the window is $V_n = h_n^d$.

This paper proposed a relaxed Parzen Window model. It tries firstly a smaller Parzen Window. If it captures the location, then the location has high degree of recommendation. If not, the model gradually relaxes the window (increases the size) and tries to capture the location and decreases the degree of recommendation of the location accordingly.

We have applied fuzzy set theory to incorporate both the distance vector and the population density. For distance vector we have taken the membership function as:

$$S_{distance}(x : \gamma, \delta) = \begin{cases} 0, & x < \gamma; \\ 2(\frac{x - \gamma}{\delta - \gamma})^2, & \gamma \leq x < \frac{\gamma + \delta}{2}; \\ 1 - 2(\frac{x - \delta}{\delta - \gamma})^2, & \frac{\gamma + \delta}{2} \leq x < \delta; \\ 1, & x \geq \delta. \end{cases}$$

For distance vector, it is assumed that the distance is moderate if it is between d_1 and d_2 where $0 < d_1 < d_2$. The smaller the distance the lesser the need. One interesting points is that anything less than d_1 is minimally required, but not every distance in reality is realistic. Here the concept of linguistic variable has been introduced. The key idea is “Distance may be less than d_1 but not very less” will be considered for minimally required. Highly required is anything $> d_2$.

Now we have to find out the value of γ and δ for distance membership functions. As we have taken distance between d_1 and d_2 as moderate, then anything between d_1 and d_2 is medium. We have introduced three linguistic variables Low, Medium and High where,

$$\text{Low is defined as: } \Phi_{low}(x : a, b) = \begin{cases} 1, & x < a; \\ 1 - 2\left(\frac{x-b}{b-a}\right)^2, & a \leq x < \frac{a+b}{2}; \\ 2\left(\frac{x-a}{b-a}\right)^2, & \frac{a+b}{2} \leq x < b; \\ 0, & x \geq b. \end{cases}$$

$$\text{Medium is defined as: } \Pi_{medium}(x : a, b) = \frac{1}{1 + \left(\frac{x-b}{b-a}\right)^2} \text{ and}$$

$$\text{High is defined as: } \Phi_{high}(x : a, b) = \begin{cases} 0, & x < a; \\ 2\left(\frac{x-a}{b-a}\right)^2, & a \leq x < \frac{a+b}{2}; \\ 1 - 2\left(\frac{x-b}{b-a}\right)^2, & \frac{a+b}{2} \leq x < b; \\ 1, & x \geq b. \end{cases}$$

clearly, mid value of $\Phi_{low}(x : a, b)$ is at d_1 , the mid value of $\Pi_{medium}(x : a, b)$ is at both d_1 and d_2 and mid value of $\Phi_{high}(x : a, b)$ is at d_2 . So, we can write,

$$\begin{aligned} \Phi_{low} : & \{ (a_{\Phi_{low}} + b_{\Phi_{low}}) / 2 = d_1 \text{ and } b_{\Phi_{low}} = (d_2 + d_1) / 2 \\ \Pi_{medium} : & \{ a_{\Pi_{medium}} - b_{\Pi_{medium}} = d_1 \text{ and } a_{\Pi_{medium}} + b_{\Pi_{medium}} = d_2 \\ S_{high} : & \{ a_{\Phi_{high}} = (d_2 + d_1) / 2 \text{ and } (a_{\Phi_{high}} + b_{\Phi_{high}}) / 2 = d_2 \end{aligned}$$

from the above mentioned equations, we can obtain the following values:

$$\begin{aligned} \Pi_{medium} : & \{ a_{\Pi_{medium}} = (d_2 + d_1) / 2 \text{ and } b_{\Pi_{medium}} = (d_2 - d_1) / 2 \\ \Phi_{high} : & \{ a_{\Phi_{high}} = (d_2 + d_1) / 2 \text{ and } b_{\Phi_{high}} = (3d_2 - d_1) / 2 \\ \Phi_{low} : & \{ a_{\Phi_{low}} = (3d_1 - d_2) / 2 \text{ and } b_{\Phi_{low}} = (d_2 + d_1) / 2 \end{aligned}$$

Now, we can say that γ and δ for distance membership function can be defined as $\gamma = a_{\Phi_{low}} = (3d_1 - d_2) / 2$ and $\delta = b_{\Phi_{high}} = (3d_2 - d_1) / 2$.

Figure 2(b) shows the membership function for distance attribute.

In case of population, we have adapted the membership function as

$$S_{population}(x : \sigma, \lambda) = \begin{cases} 0, & x < \sigma; \\ 2\left(\frac{x-\sigma}{\lambda-\sigma}\right)^2, & \sigma \leq x < \frac{\sigma+\lambda}{2}; \\ 1 - 2\left(\frac{x-\lambda}{\lambda-\sigma}\right)^2, & \frac{\sigma+\lambda}{2} \leq x < \lambda; \\ 1, & x \geq \lambda. \end{cases}$$

and if it is given that the population is moderate between p_1 and p_2 , then in the same way- discussed above for distance vector- we can adjust the value of σ and λ and we may write $\sigma = (3p_1 - p_2) / 2$ and $\lambda = (3p_2 - p_1) / 2$.

We define the fuzzy conjunction of two membership values $\mu_A(\cdot)$ and $\mu_B(\cdot)$ as Hamacher product which can be defined as:

$$\text{Hamacher}(\mu_A(\cdot), \mu_B(\cdot)) = \frac{\mu_A(\cdot) \times \mu_B(\cdot)}{\mu_A(\cdot) + \mu_B(\cdot) - [\mu_A(\cdot) \times \mu_B(\cdot)]}$$

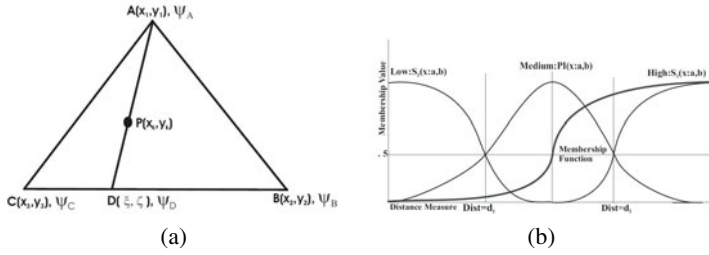


Fig. 2. (a) The nearest neighboring population points of suggested point. [The population points are the corners of the surrounding triangle of the suggestion point]. (b) The membership function for distance attribute.

and the pattern volume is defined as $1 - Hamacher(\mu_{distance}(\cdot), \mu_{population}(\cdot))$ because the greater the Hamacher value, the better the matching. So, through a smaller window it can be viewed. So, we have taken complement of Hamacher product as the volume. Algorithm 4 describes the technique in a precise way.

Algorithm 4. Relaxed Parzen Window based Fuzzy Multi-objective Model

```

Input: Point  $p(x, y)$  ▷ For which the need is to be calculated
1:  $x_d \leftarrow \mu_{distance}(p(x, y))$ 
2:  $x_p \leftarrow \mu_{population}(p(x, y))$ 
3:  $PatternVolume \leftarrow 1 - Hamacher(x_d, x_p)$  ▷ The conjunction of  $x_d$  and  $x_p$ 
4:  $WindowVolume \leftarrow V_{small}$  ▷ Initialize the Parzen Window volume
5: if ISVISIBLETHROUGHPARZENWINDOW( $PatternVolume$ ) = True then
6:   return HighRequirement
7: else
8:   RELAX( $WindowVolume$ )
9:   if ISVISIBLETHROUGHPARZENWINDOW( $PatternVolume$ ) = True then
10:    return MediumRequirement
11:   else
12:    RELAX( $WindowVolume$ )
13:    if ISVISIBLETHROUGHPARZENWINDOW( $PatternVolume$ ) = True then
14:     return LowRequirement
15:    else
16:     return NoRequirement
17:    end if
18:   end if
19: end if

```

4 Result and Discussion

Some input data and model-suggested output is given in table 1 and table 2 respectively. In table 1 the location of the facility centers, the boundary lines of the map and the

population points with population value are given. The boundary lines are given by its two end points. The population of a point means the population of the Block of the District.

Table 1. Input Data Set (Facility Centers, Boundary Lines & Population)

Facility Centers	Boundary Lines		Population Information	
	Start Locations	End Point	Population Point	Population
(66, 66); (111, 111); (144, 144);	(401, 321)	(377, 305)	(58, 58)	156320
(179, 179); (132, 132); (93, 93);	(377, 305)	(352, 281)	(106, 62)	110393
(41, 41); (175, 175); (168, 168);	(352, 281)	(330, 266)	(159, 146)	212742
(65, 65); (219, 219); (111, 111);	(330, 266)	(309, 241)	(145, 107)	242377
(74, 74); (376, 376); (188, 188);	(309, 241)	(282, 227)	(39, 91)	289903
(194, 194); (79, 79); (222, 222);	(282, 227)	(258, 208)	(92, 119)	475439
⋮	⋮	⋮	⋮	⋮

Table 2. The Fuzzy Decision Making Output Data

	Projected Location	Distance Vector	Projected Population	Fuzzy Membership Values		Hamacher Product	Final Suggestion
				Distance	Population		
1	(622, 224)	92	242964	1.0	1.0	1.0	High
2	(452, 260)	76	179860	0.930	0.874	0.820	High
3	(421, 338)	84	162379	0.999	0.771	0.771	Moderate
4	(370, 192)	75	180731	0.919	0.878	0.815	High
5	(677, 342)	57	110374	0.291	0.323	0.180	No
6	(511, 376)	59	192383	0.372	0.929	0.362	Low
	⋮	⋮	⋮	⋮	⋮	⋮	⋮

In table 2, some of the output data are given. In row 1 and 2 of table 2, both distance and population is high. So, the decision is “High”. Likewise, one “High” and one “Moderate” contribution may give suggestion “Medium” as in case of row 3 of table 2. “Low” recommendation is because one of the factors is “Low” or both the factors are “Moderate” as in row 6 of table 2. The recommendation is “No” if neither the distance is far nor the population is dense as in row 5 of table 2.

The time complexity of the addition of a point to the triangle set using this algorithm is $O(\sqrt{n})$ where n is the number of existing cites and that of Bowyer-Watson algorithm is $O(n^{1.5})$. The time complexity of the model i.e. the time complexity of Algorithm 3 is $O(n^{1.5} + t^{1.5} + |B|)$ for new facility point entry where $|B|$ is the number of boundary lines, n is the existing facility points and t is the known population points. The final graphical output is given in figure 3.

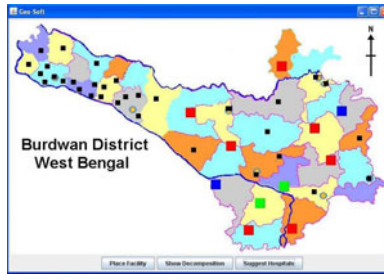


Fig. 3. The final output of the system. The Red color points are highly recommended, Blue are moderate and the Green are minimally recommended suggestion points.

5 Conclusion

The paper suggested a Relaxed Parzen Window based Fuzzy Decision making model that generates graphical output in GIS based platform. The data have been collected from the District Statistical Handbook of Burdwan-2004 of Bureau of Applied Economics and Statics, Govt. of West Bengal, India. The model can manage a high degree of complexity in different situations and shows a very good suggestion for possible facility locations. As the model considers both population and distance vector as parameters it is more rational and realistic. There are enough scope of development of the model using Soft computing and Decision Theoretic approach.

References

1. Berg, M.D., Cheong, O., van Kreveld, M., Overmars, M.: Computational Geometry- Algorithms and Applications, 3rd edn. Springer, Heidelberg
2. Bowyer, A.: Computing Dirichlet tessalations. *The Computer Journal* 24(2), 162–166 (1981)
3. Chew, L.P., <http://www.cs.cornell.edu/home/chew/chew.html> (access April 24, 2011)
4. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, ch. 4, 2nd edn., pp. 161–177. John Wiley (2008)
5. Fleming, G., van der Merwe, M., McFerren, G.: Fuzzy expert systems and GIS for cholera health risk prediction in Southern Africa. *ELSEVIER-Environmental Modelling and Software* 22, 442–448 (2007)
6. Gold, C.: <http://www.voronoi.com/> (access March 27, 2011)
7. Gold, C.: Problems with handling spatial data-the Voronoi approach. *Cism Journal Acsgc* 45(1), 65–80 (1991)
8. Lo, C.P., Yeung, A.K.: Concepts and Techniques of Geographic Information Systems, 2nd edn., Prentice Hall of India
9. Watson, D.F.: Computing the n-dimensional Delaunay Tessalation with application to Voronoi polytops. *The Computer Journal* 24(2), 167–172 (1981)
10. Yang, X., Cui, W.: A novel spatial clustering algorithm based on delaunay triangulation. *Journal of Software Engineering and Applications* 3, 141–149 (2010)
11. Yen, J., Langari, R.: Fuzzy Logic-Intelligence, Control, and Information. Pearson Education

Review on Cost Effective Software Engineering Using Program Slicing Techniques

S. Koushik and R. Selvarani

Department of Computer Science,
M S Ramaiah Institute of Technology,
Bangalore - 560054
{koushik.msrit,selvarani.riic}@gmail.com

Abstract. Software Development is a complex and multidimensional task. The development process should not only concentrate on just writing the code but also focus on the cost effective methods. Cost and time are the major constraints of software development process. Program slicing has unique importance in addressing the issues of cost. It is a program analysis technique which provides mechanism to analyze and understand the program behavior for further restructuring and refinement. Many people have investigated the relationship between program slicing and software development phases on the basis of empirical studies conducted in the past and also establish the fact that how program slicing can be helpful in making software system cost and time effective. This paper provides a general overview of program slicing which is a cost effective software engineering technique.

Keywords: Software Engineering, Cost Effective, Program Slicing, Cost Estimation.

1 Introduction

Program slicing as the name describes itself slicing a program into several important chunks of code. Previously different approaches adopted in optimization and analysis of program code. But program slicing is far better choice than conventional techniques in terms of saving cost and time. This paper consists of analysis of program slicing effectiveness on the basis of empirical studies conducted in past [4], experimental results [5] and theoretical conclusions.

Program testing is one of the important phases in software development [7]. Testing can also be considered as investment to find out bugs. Software testing is an investigation conducted to provide stakeholders with information about the quality of the product or service under test. Software testing also provides an objective, independent view of the software to allow the business to appreciate and understand the risks of software implementation [8]. Test techniques include, but are not limited to, the process of executing a program or application with the intent of finding software bugs (errors or other defects). There are mainly two types of testing viz. manual testing and automated testing.

This paper is further organized as follow. Section 2 briefly explain overview of the program slicing and few techniques, Section 3 describes debugging and program analysis application of program slicing in software debugging, Section 4 evaluates the simultaneous dynamic slicing and the behavior of simultaneous dynamic slicing under different parameters.

2 Program Slicing Overview

One of the most efficient techniques of solving a problem is to break down a huge problem to smaller components [1, 10]. Similarly, breaking down a large code or project into several smaller components makes the task of software development easier and more cost effective [1]. Breaking code into smaller components include debugging, testing, program comprehension, restructuring, downsizing, and parallelization. [3].

Simplifying a program by eliminating the parts of the program which is not affecting the actual working of the program, this technique is called program slicing. Program slicing was first introduced by Mark Weiser [1]. The main aim of program slicing is to identify and extract relevant parts of a software program from a more complicated code. A slice is a small piece of the original program which preserves the actual behavior. Slicing criterion consists of pair (line no, variable). This slicing criterion is used to create slices in the program. Program slicing, which is based on internals of the code, can be most obviously applied to structural testing techniques.

2.1 Static Slicing

Slicing criterion decides the slice of any program. Now let us see the definition of slicing criterion. Weiser [1] developed the slicing technique, a slicing criterion is defined as a pair values $\langle p, V \rangle$, and where p is a program point and V is a subset of program variables.

This $\langle p, V \rangle$ pair is a subset of program statements that preserves the actual behavior of the original program. The program point p is the point of program where the slicing point is set with respect to the program variables in V . The actual behavior of the program has to be preserved even after slicing on any input, which is in a way static. Hence the name static slicing was given to this type of slicing [1].

According to the above definition, an algorithm to compute slices has been proposed which by Mark Weiser [1] uses the method of backward traversals of the Program Dependence Graph (PDG).

2.2 Dynamic Slicing

Another method of slicing is dynamic slicing. The main disadvantage of a static slice is that it may very often contain statements which have no influence on the values of the variables of interest. This particular behavior for certain programs also led to some kind of anomalous behavior in the program.

To over come this behavior Korel and Lasky [9] proposed an alternative slicing definition, namely dynamic slicing. This dynamic slicing uses dynamic analysis to

identify all and only the statements that affect the variables of interest on the particular anomalous execution trace.

This method also helps in reducing the size of the slice to a greater extent, thus allowing an easier localization of the bugs. Arrays and pointers can be handled in very efficiently using dynamic slicing technique. Each array is considered as an array element in dynamic slicing. Also, dynamic slicing distinguishes which objects are pointed to by pointer variables during a program execution.

2.3 Quasi Static Slicing

We might not require dynamic slicing all the time. Sometimes we also require the static slicing criterion to get the best use of the slicing technique [2]. This gave rise to a technology called Quasi Static Slicing. This was the first attempt to define a hybrid slicing method ranging between static and dynamic slicing [3] and [4]. There is a need to check the behavior of the application for some fixed input variables as well as some inputs which vary at runtime. This requirement led to a hybrid model called quasi static slicing. As we know already have explained that the slicing behavior should be capable of preserving the behavior of the original program with respect to the variables of the slice criteria as well as concentrate on the possible program inputs as well, quasi static slicing preserves both the criterion while creating slices. The subset of inputs is specified by the possible combination of values that the unconstrained input variables might assume [15].

The quasi static slice coincides with both static and dynamic slice where in static slice the values of all input variables are fixed, and the slice is a dynamic slice [14]. The quasi static slicing is closely related to partial evaluation or mixed computation [5]. Mixed computation is a technique to specialize programs with respect to partial inputs. Slices in quasi static slices are computed on specialized programs [3].

2.4 Simultaneous Dynamic Slicing

Simultaneous dynamic slicing calculates slices with respect to a set of program executions. This method of slicing was introduced by Hall [10]. This type of dynamic slicing applies the slicing criterion to a set of test cases rather than just one test case, hence the name simultaneous dynamic slicing [9].

A simultaneous program slice on a set of test cases is not simply given by the union of the dynamic slices on the component test cases. Indeed, simply union of dynamic slices is unsound, in that the union does not maintain simultaneous correctness on all the inputs. Therefore, Hall [10] proposed an iterative algorithm that, starting from an initial set of statements, incrementally builds the simultaneous dynamic slice. The value of slicing criteria is computed at every iteration a larger dynamic slice. Simultaneous dynamic slicing has been used to locate functionality in code. The set of test cases can be seen as a kind of specification of the functionality to be identified.

2.5 Conditioned Slicing

Conditioned slicing is a technique used to compute program slices with respect to a subset of program executions [14]. Conditioned slice is an extension of static slicing

which consists of $\langle V, n, F \rangle$ pair. The above pair is defined as V is a set of variables, n is a point in the program, and F is a condition for slice. A conditioned slice consists of a subset of program statements which preserves the behavior of the original program with respect to a slicing criterion for any set of program executions [9]. Conditioned slicing allows a better decomposition of the program giving programmer the ability to analyze code slices with respect to different perspectives.

Canfora et al [11] have demonstrated that conditioned slicing subsumes any other form of statement deletion based slicing method, i.e., the conditioned slicing criterion can be specified to obtain any form of slice.

2.6 Comparison of Different Slicing Techniques

The above table summarizes the different slicing technique for a particular program with 14,500 lines of code and 183 function points. Distance, slice size and percentage affect gives the analysis of different slicing techniques on the program. These points are observed with the help of above definitions and David et. al. [14] demonstrations.

Table 1. Slicing Techniques analyzed with different parameter

	LOC comments and lines	Without and blank	Function Points	Distance	Slice Size	Percentage affect
Static slicing	14,500		183	110	194	19%
Dynamic slicing	14,500		183	120	196	36%
Quasi static slicing	14,500		183	97	202	25%
Simultaneous dynamic slicing	14,500		183	131	199	42%
Conditioned slicing	14,500		183	111	194	11%

3 Debugging and Program Analysis

Increasing the quality and productivity of software development and its processes is an important research objective of software engineering. It is widely recognized and accepted that errors have a large impact on software productivity and quality. These errors in the program are called bugs sometimes [13].

Debugging is a process or a method of finding and reducing bugs, defects or errors, in a software program or hardware, achieving expected behavior. Debugging is a tedious task when various subprograms are tightly coupled. Debugging a subprogram might lead to cause a bug in another subprogram as a result. Debugging is one of the major concerns in any software development life cycle. It is vary difficult to localize and identify faults and debug very large and complex systems, making debugging process more difficult than expected [6].

Program slicing is very efficient and promising approach to localize faults efficiently [7]. Debugging concentrates on a particular module or subsystem for localizing faults and identifying bugs. There are several steps one has to follow while debugging a program. The general steps for debugging are as described below. First

step in debugging is to identify external symptoms of the software and translating and relating these symptoms to internal symptoms which represent the problem in terms of data and control flow problems in the program. After identifying the symptoms, creating and applying a slicing method and related slicing criterion on the program based on the selected data or control flow. Using this criterion, a slice is obtained containing code that caused failure [18]. Once the slice is obtained, a program state is found to check the previous state and the program state is restored to a point when the control last reached the program step. The final step is to observe the values of some variables in the restored state to find fault in the program. If search is not successful then user may choose to further examine the restored state, guess a new fault, or select a new slicing criterion and repeat the cycle until the fault is localized.

Let us see few experimental results and analysis conducted by Kusumutu et al [8].

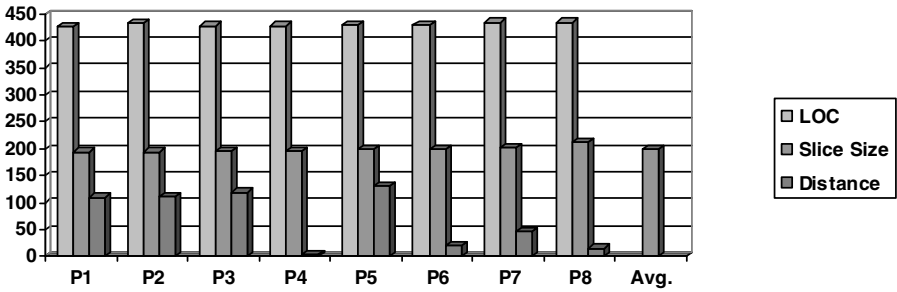


Fig. 1. Size of program and slice

Figure 1 shows the size of the program and of the slice obtained by the typical slicing criterion for failure of the program.

Figure 2 and 3 show the data about the time (in minutes) required to localize each of the faults in each trial using without slicing and with slicing.

The critical phase of software development life cycle is software testing [17]. Software testing involves large human resource to be dedicated. Hence we discuss the application of program slicing technique in the rest of the chapters.

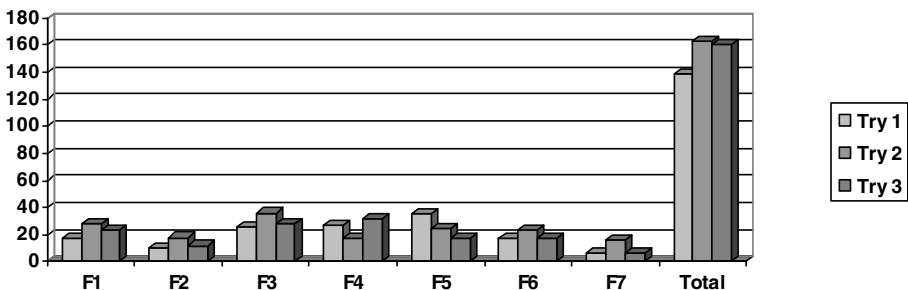


Fig. 2. Fault localization results with slicing

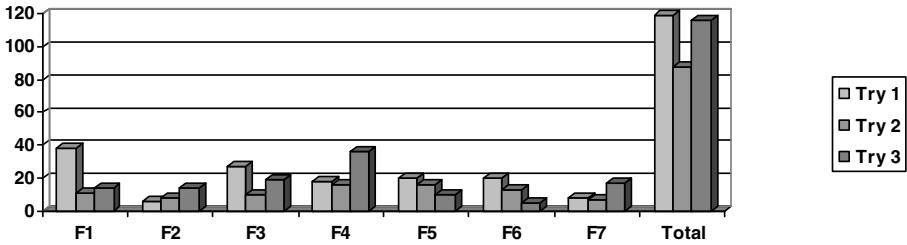


Fig. 3. Fault localization results without slicing

4 Overview of Simultaneous Dynamic Slicing Technique

Depending on the above discussions and definitions, it is observed that simultaneous dynamic slicing is a better choice [18]. Analysis of slicing techniques can be established with theoretical facts or by experimental results. In this section we try to establish that simultaneous dynamic slicing technique is a better choice for slicing, based on the theoretical fact which is also called as empirical analysis or studies of processes.

Analysis of program slicing can be done using various parameters. A few parameters can be considered to analyze the simultaneous dynamic program slicing like Lines of Code, Slice size, Distance, functional points, variables, execution time, and type of output. According to the definitions given in the previous chapters we can say that simultaneous dynamic slicing is a better technique compared to all the others mentioned [10].

Subset of inputs is chosen instead of full set of inputs in simultaneous dynamic slicing [8]. This reduces the time of execution which in turn will affect on cost. In simultaneous dynamic slicing the slicing criteria is to apply to a subset of inputs instead of all the inputs. Let us define the above criterion and see how these factors affect the dynamic program slicing. “Lines of code” is the actual number of lines the application or the programs have. Slice size is the size of each block of code after the slicing criteria is applied [19]. Distance is the total distance the slice can travel effectively to make the program execution more efficient. Functional points are the number of functions written for an application, this functional point also helps in determining the cost of the whole project. Variables are elements which are viable to vary at any point of time [10]. The program execution and slicing also depends on the variables used as more number of variables can result in delayed execution time. Execution time is the actual time taken to execute all the functions of the particular program. Simultaneous dynamic slicing executes best under all the conditions given above as the definition of the simultaneous dynamic slicing is stated as a program slice which is decided on a set of executions. As the slice size is dynamically calculated each iteration the lines of code doesn’t affect the slice criteria. Applying of slice is dynamic the slice size is calculated on runtime which saves time in calculating the slice size. As the simultaneous dynamic slice is based on a set of inputs and a subset of functions the distance covered will be large based on the slice. This also covers as many functions points as possible. Variables which are sometimes local to a function come within the slice which creates a subset of variables making faster execution.

This technique of slicing has one drawback. The execution time when compared to static slicing. In static slicing the slice size is decided before the execution. In simultaneous dynamic slicing, slices are calculated dynamically during the runtime. Each iteration starts with a new slice that is calculated. This takes more time in creating the slices [9].

5 Conclusion

The prime objective of this paper is to analyze the previous program slicing techniques developed by different researchers. This paper establishes the fact that program slicing mechanism can play a vital role in program testing for building robust and efficient programs. In future, program slicing technique in the phase of testing can be standardized. This slicing technique also reduces the cost of the management.

References

- [1] Weisier, M.: Program Slicing. *IEEE Transaction on Software Engineering*, 439–449 (1984)
- [2] Xu, B., Xi, J.Q., Zhang, X.: A Brief Survey of Program Slicing. *ACMSIGSOFT Software Engineering Notes*, 1–36 (2005)
- [3] Binkley, D., Harman, M.: Empirical Study of Optimization Techniques. *Loyola College in Maryland*, 1–24 (1997)
- [4] Cetinkaya, O., Cetinkaya, D.: An Empirical Study of Static Program Slice size. *ACM Transaction on Software Engineering and Methodology (TOSEM)*
- [5] Ishio, T., Kusumoto, S.: Program Slicing Tool for Effective Software Evolution Using Aspect-Oriented Technique. In: *Proceedings of the 6th International workshop on Principles of Software Evolution*, pp. 49–76 (2003)
- [6] Binkely, D.: Application of Program Slicing to Regression Testing. *Information and Software Technology Issue on Program Slicing* 16(2) (1999)
- [7] Fujioka, A., Okamoto, T., Ohta, K.: *Investing in Software Testing*, Australia, pp. 244–251 (1992)
- [8] Kusumoto, S., Akira: Experimental Evaluation of Program Slicing for Fault Localization. *Empirical Software Engineering* 7, 49–76 (2002)
- [9] Korel, B., Laski, J.: Dynamic slicing of computer programs. *The Journal of Systems and Software* 13(3), 187–195 (1990)
- [10] Hall, R.J.: Automatic extraction of executable program subsets by simultaneous program slicing. *Journal of Automated Software Engineering* 2(1), 33–53 (1995)
- [11] Canfora, G., Cimitile, A., De Lucia, A.: Conditioned program slicing. *Information and Software Technology* 40(11/12), 595–607 (1998)
- [12] Nielson, F., Nielson, H.R., Hankin, C.: *Principles of program analysis* (2005)
- [13] Kusumoto, S., Nishimatsu, A., Nishie, K., Inoue, K.: *Experimental Evaluation of program slicing for fault localization* (2002)
- [14] Binkley, D., Harman, M.: *A Survey of Empirical Results on Program Slicing* (2003)
- [15] Tip, F.: *A Survey of Program Slicing Techniques* (2006)
- [16] Mohapatra, D.P., Mall, R., Kumar, R.: *An Overview of Slicing Techniques for Object-Oriented Programs* (2006)
- [17] Hierons, R.M., Harman, M.: *Program Analysis and Test Hypotheses Complement* (2010)
- [18] De Lucia, A.: *Program Slicing: Methods and Applications*, pp. 1–8 (2003)

Binarization of Document Images Using Hierarchical Histogram Equalization Technique with Linearly Merged Membership Function

Satadal Saha¹, Subhadip Basu², and Mita Nasipuri²

¹ MCKV Institute of Engineering, Liluah, Howrah, India

² Jadavpur University, Kolkata, India

{satadalsaha, subhadip}@ieee.org,

mmasipuri@cse.jdvu.ac.in

Abstract. Binarization itself is a process of finding a threshold value for converting a grey level image into a binary image. The threshold may vary depending on whether it is found globally or locally. It is found that either of the global and the local threshold itself can not provide a good binarization; rather a combination of the two is a better solution. In the current work, we have applied histogram equalization technique over the complete image and also over all the partitions of the image at different levels of hierarchy. A novel scheme is formulated for giving the membership value to each pixel at each level of hierarchy during histogram equalization. Then the image is binarized depending on the net membership value of each pixel. The technique outperforms when exhaustively tested on document images collected from different sources.

1 Introduction

Binarization is the method of converting a grey scale image (popularly known as multi-tone image) into a black-and-white image (popularly known as two-tone image). This conversion is based on finding a threshold grey value and deciding whether a pixel having a particular grey value is to be converted to black or white. Usually within an image the pixels having grey value greater than the threshold is transformed to white and the pixels having grey value lesser than the threshold is transformed to black at the time of binarization. Binarization has been the area of research for last twenty years or so, mainly to find a single threshold value or a set of threshold values for converting a grey scale image into a binary image. Most of the algorithms till developed are of generic type with or without using local information or special content within the image.

The most convenient and primitive method is to find a global threshold for the whole image and binarize the image using the single threshold. In this technique, the local variations are actually suppressed or lost, though they may have important contribution towards the information content within it. On the other hand, in case of determining the threshold locally, a window is used around a pixel and threshold value is calculated for the window. Now depending on whether the threshold is to be

used for the center pixel of the window or for all the pixels in the window, the binarization is done on pixel-by-pixel basis, where each pixel may have a calculated threshold value, or on region-by-region basis where all pixels in a region or window have same threshold value.

The major contribution of research for binarization is to recover or extract information from a degraded document images. Otsu [1] developed a method based on grey level histogram and it maximizes the intra-class variance to total variance. Sauvola [2] developed an algorithm for text and picture segmentation within an image and binarized the image using local threshold. Gatos [3] used Wiener filter and Sauvola's adaptive binarization method. Valverde [4] binarized the image using Niblack's technique. A slight modification of Niblack's method is done in [5] by Zhang. Milewski [6] used a binarization based method for separating handwritten text from carbon copy medical form images. A comparative study of different document image binarization methods have been discussed in [7]. In our earlier work [8], a histogram equalization based technique has been used hierarchically for camera captured image binarization.

The main objective of the present work is to attain a balance between the global threshold and the local threshold such that the effective binarization enriches our earlier work [8]. We have applied histogram equalization method over the whole image globally and also over different image partitions at each hierarchical level. A scheme is formulated for giving the membership value to each pixel at each level of histogram equalization. The net membership value of each pixel is then calculated by linearly merging the membership values and the image is then binarized depending on the net membership value of each pixel.

2 Present Work

The grey value variation within an image gives the impression of contrast in it. A high grey value variation within an image always provides a high contrast and in turn a highly contrasted image facilitates the decision of making a grey level pixel either to black or to white during the time of binarization. In the current scheme, histogram equalization is done over the whole image as well as over the different area of localization. Localization is done by dividing the image by two both horizontally and vertically thereby generating four quadrants. Histogram equalization at each level of localization provides a membership of greyness for each pixel. Ultimately all the membership values thus obtained for different levels are combined to get the net membership value of greyness for each pixel. Each pixel is then binarized depending on whether the net membership values crosses 0.5 or not. If it crosses 0.5 then it is made white, otherwise it is made black.

2.1 Dataset Generation

A dataset is generated over which the technique of binarization is tested and the performance is evaluated. The dataset consists of both grey and color images collected from different sources, as given below:

- News paper pages – 115 images
- Book pages – 124 images
- Magazine pages - 82 images
- Surveillance camera captured scenes – 245 images
- Handwritten document pages – 144 images

Over and above, the sample document images uploaded in web in DIBCO 2009 (4 images) and H-DIBCO 2010 (10 images) competitions are also used as benchmark for performance measure of the technique. As a whole the dataset consists of 724 images.

2.2 Pre-processing

Images gathered from different sources are impaired by different noises. And also the whole technique is applied over the grey scale version of the original image. Following preprocessing techniques are implemented in the current work to address the issues mentioned above.

Grey scale conversion: For each pixel, 24-bit color value is converted to 8-bit grey value using the formula [9] written in equation (1).

$$grey(i, j) = 0.59 * R(i, j) + 0.30 * G(i, j) + 0.11 * B(i, j) \quad (1)$$

where, (i, j) is the position of a pixel in the image and $grey(i, j)$ ranges from 0 to 255.

Median filtering: As the proposed technique is based on detecting the edge of the image, salt-and-peeper noise generated randomly needs to be removed from the image. Median filter is a non-linear filter which removes salt-and-peeper noise from the image. This filter replaces the grey value of a pixel by the median of the grey values of its neighbors [9]. In this work, a 3×3 mask is used as median filter

2.3 Histogram Equalization

Contrast of each image is enhanced through histogram equalization technique, as discussed in [9]. Total 256 numbers of grey values (0 to 255) are used for stretching the contrast. Let the total number of pixels in the image be N and the number of pixels having grey value k be n_k . Then the probability of occurrence of grey value k is, $P_k = n_k / N$. The stretched grey value S_k is calculated using the cumulative frequency of occurrence of the grey value k in the original image using the formula:

$$S_k = \sum_{j=0}^k \frac{n_j}{N} \times 255 \quad (2)$$

where, 255 indicates the maximum grey value in the enhanced image. This S_k divided by 255 results an enhanced fractional grey value ($f_g = S_k / 255$) of a pixel in the range 0 to 1 and resembles the likeliness of a pixel to be white. A fraction close to 1 indicates the pixel to be white. On the other hand a fraction close to 0 indicates the pixel to be black. Application of histogram equalization method over the whole image gives 0th level membership value ($f_{mem}(i, j, 0) = f_g$) of pixel (i, j) .

2.4 Hierarchical Partitioning of Image and Application of Histogram Equalization in Each Image Partitions

By partitioning the image midway, both width wise and height wise, four equal segments of the image are generated and histogram equalization method is applied in each segment separately using the original grey value of the image. This provides 1st level membership value, $f_{mem}(i,j,1)$ of pixel (i,j) . Likewise each of the 1st level segments is divided into four equal quadrants and histogram equalization method is applied over each quadrant to get the 2nd level membership value, $f_{mem}(i,j,2)$ of pixel (i,j) . This process goes on up to nth level and membership value $f_{mem}(i,j,n)$ is obtained for pixel (i,j) . Fig. 1 shows the partitioning of the image from 1st level to 4th level, excluding the 0th level.

<p>टन की तुलना में बहुत कम था। पिछले सप्ताह से लागू यह प्रतिबंध 31 दिसंबर तक जारी रहेगा। गुजरात के किसान फिलहाल गेहूँ की कटाई कर रहे हैं और रोजाना 500-800 टन का आवक का अनुमान है। मुंबई के अंतरराष्ट्रीय कारोबारी घराने के अनुसार अपुष्ट रिपोर्ट से पता चलता है कि पिछले माह गुजरात</p>	<p>टन की तुलना में बहुत कम था। पिछले सप्ताह से लागू यह प्रतिबंध 31 दिसंबर तक जारी रहेगा। गुजरात के किसान फिलहाल गेहूँ की कटाई कर रहे हैं और रोजाना 500-800 टन का आवक का अनुमान है। मुंबई के अंतरराष्ट्रीय कारोबारी घराने के अनुसार अपुष्ट रिपोर्ट से पता चलता है कि पिछले माह गुजरात</p>
(a)	(b)
<p>टन की तुलना में बहुत कम था। पिछले सप्ताह से लागू यह प्रतिबंध 31 दिसंबर तक जारी रहेगा। गुजरात के किसान फिलहाल गेहूँ की कटाई कर रहे हैं और रोजाना 500-800 टन का आवक का अनुमान है। मुंबई के अंतरराष्ट्रीय कारोबारी घराने के अनुसार अपुष्ट रिपोर्ट से पता चलता है कि पिछले माह गुजरात</p>	<p>टन की तुलना में बहुत कम था। पिछले सप्ताह से लागू यह प्रतिबंध 31 दिसंबर तक जारी रहेगा। गुजरात के किसान फिलहाल गेहूँ की कटाई कर रहे हैं और रोजाना 500-800 टन का आवक का अनुमान है। मुंबई के अंतरराष्ट्रीय कारोबारी घराने के अनुसार अपुष्ट रिपोर्ट से पता चलता है कि पिछले माह गुजरात</p>
(c)	(d)

Fig. 1. Hierarchical partitioning of the image at (a) 1st level, (b) 2nd level, (c) 3rd level, (d) 4th level

2.5 Formulation of Net Membership Value of Each Pixel

As discussed in section 2.4, for each level of contrast enhancement through histogram equalization, each pixel gets an enhanced fractional grey value leading to a membership value, $f_{mem}(i,j,level)$ indicating that whether the said pixel is closer to white or black. These membership values for each pixel obtained at different hierarchical levels are combined to get a net membership value of the pixel. Now assuming that the global scenario has a little effect over the local variations, during the process of combination of the membership values, the local membership values are given more weight over global membership values. Keeping this in mind the net membership value, $f_{net_mem}(i,j,level)$ for pixel (i,j) is calculated using equation (3).

$$f_{net_mem}(i, j) = \frac{\sum_{k=startlevel}^{endlevel} f_{mem}(i, j, k) \times (k + 1)}{\sum_{k=startlevel}^{endlevel} (k + 1)} \tag{3}$$

where, the summation (Σ) is over the levels *startlevel* to *endlevel*. Thus during the process, each pixel gets a net membership value in which it retains the global as well as the local information of the variation of grey level. It is also to be observed that the membership value at a level *k* is multiplied by (*k+1*). This is because the starting *level* is considered as 0. In this way the membership values obtained at more depth in the hierarchy, i.e. at more localized region, are given more weight during the calculation of net membership value.

2.6 Binarization

As discussed in the last two subsections, for each pixel a net membership value is obtained, which is compared with the threshold value for binarization. As the image is histogram equalized, the mean grey level lies in the middle position of the grey scale range (0 to 255). So 0.5 is chosen as threshold and any pixel having net membership value greater than 0.5 is converted to white; otherwise it is converted to black.

2.7 Cleaning the Image

The binarized image thus obtained contains some noise pixels here and there. These noises are removed using morphological open-close method [9]. The opening of a set A by structuring element B is defined as

$$A \circ B = \{A \ominus B\} \oplus B \tag{4}$$

So it is the erosion of A by B, followed by dilation of the result by B. The closing of a set A by structuring element B is defined as

$$A \bullet B = \{A \oplus B\} \ominus B \tag{5}$$

So it is the dilation of A by B, followed by erosion of the result by B.

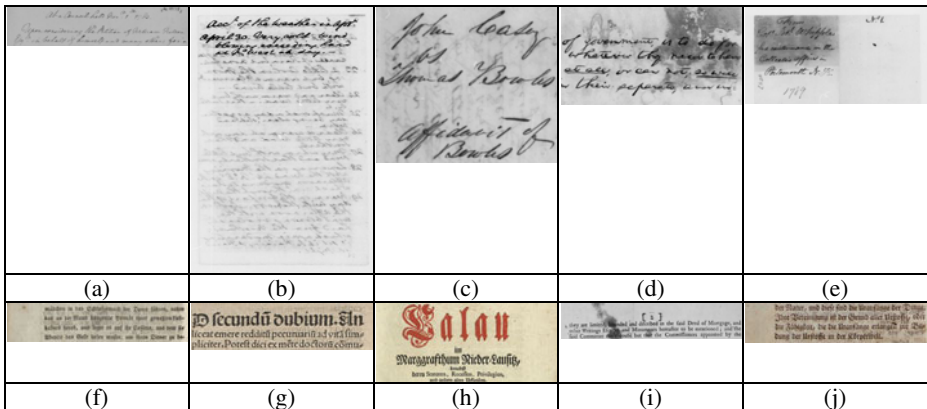


Fig. 2. Original sample images published in H-DIBCO 2010 competition

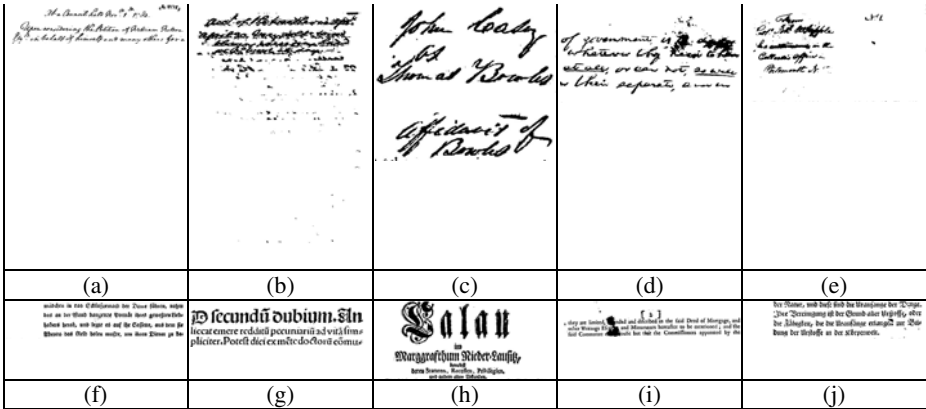


Fig. 3. Binarized version of the images shown in Fig. 2. using the proposed method

3 Experimental Results

The dataset used for the current work consists of 724 images collected from different sources, as discussed in section 2.1. Otsu’s method of binarization is used as a benchmark for comparing the performance of the proposed technique. The images are binarized using both the proposed technique and Otsu’s method of binarization.

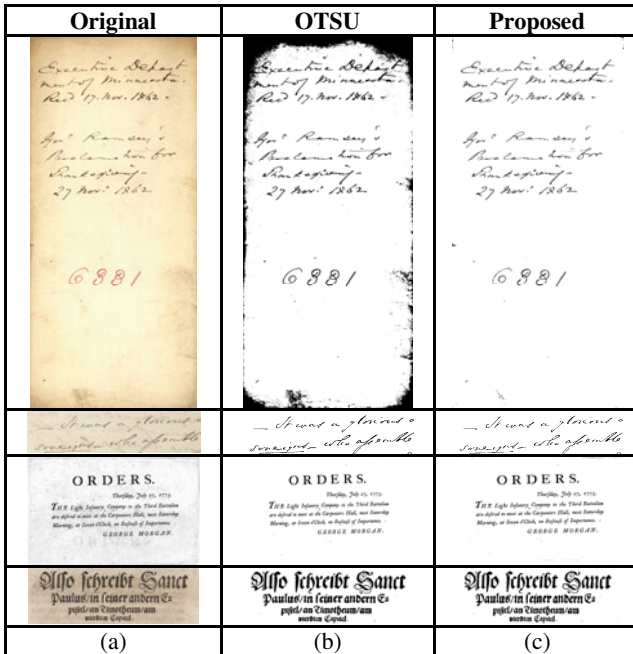


Fig. 4. Binarized version of DIBCO 2009 sample images. (a) Original images, (b) Binarized images using Otsu’s method, (c) Binarized images using proposed method.

Selecting the number of level or depth is an issue here for proper binarization. In the current work, the number of divisions and condition for application of histogram equalization depend on two factors:

- *Size of a partition*: During the process of partitioning, the height and width of each partition is divided by 2, provided each having a value of greater than 2. If either of height or width reaches that limit then its value is freezed and the other's value is divided by 2 for subsequent level of partitioning. When both height and width reach the limit, further division of sub-images is stopped.
- *Standard deviation of grey values of pixels in each partition*: For any partition at any level, the standard deviation of the grey values of the pixels in that partition is calculated. If that value is very low compared to the value at 0th level then that partition is not histogram equalized. Rather, if the grey value of a pixel in that partition is less than the mean grey value of that partition then its membership value is given as 0.4 (i.e. forcing it towards black) otherwise its membership value is given as 0.8 (i.e. forcing it towards white).

Fig. 2 shows some document images and Fig. 3 shows the binarized version of them using the proposed technique. Fig. 4 shows some images with binarized images obtained using Otsu's method and the proposed method.

4 Performance Testing and Conclusion

4.1 Comparing Output Image with the Ground Truth Image

The ground truth image corresponding to the source image contains the actual or correct binary information. For evaluating the performance of the proposed method of binarization, the ground truth image and the output binarized images are read and both the binary data are extracted. These two data are compared in point-by-point basis to get the result of matching, which can be classified into three cases:

- (a) *True positive (tp)* case: corresponding pixels in both the images are ON (black).
- (b) *False positive (fp)* case: the pixel in output image is ON but the corresponding pixel in ground truth image is OFF (white).
- (c) *False negative (fn)* case: the pixel in output image is OFF but the corresponding pixel in ground truth image is ON.

During comparison, tp , fp and fn are counted according to the three cases respectively.

4.2 Computing Recall, Precision and f-Measure

Recall (rc) is the fraction of ON pixels in the ground truth image that are ON in the output image also. *Precision (pr)* is the fraction of ON pixels in the output image that are either ON or OFF in the ground truth image. *F-measure (fm)* is a quantity combining the recall and precision, giving the performance of binarization.

$$\begin{aligned}
 rc &= \frac{tp}{fn + tp} \\
 pr &= \frac{tp}{fp + tp} \\
 fm &= \frac{2.rc.pr}{rc + pr} \times 100\%
 \end{aligned}
 \tag{4}$$

The performance of the proposed technique is evaluated by categorizing the images into three types: printed document, hand written document and camera captured frame. We have applied Otsu's binarization algorithm over the images but our method seems to be more effective in binarizing the image. The summary of the performance evaluation is given in structural form in Table 1.

Table 1. Performance summary of the proposed method as compared to Otsu's method

Document Type	%fm using Otsu	%fm using proposed method
Printed document	89.12	92.36
Hand written document	88.23	91.45
Camera captured scene	85.32	88.22
Average	87.56	90.68

As seen from Table 1, the proposed technique is well suited in binarizing different types of images. The proposed technique not only performs better than Otsu's method, but also it improves the efficiency of our earlier work [7] to a greater extent.

Acknowledgments. Authors are thankful to the CMATER, the SRUVM project, and the PURSE program of C.S.E. Department, Jadavpur University, for providing necessary infrastructural facilities during the progress of the work. Mr. S. Saha is thankful to the authorities of MCKV Institute of Engineering for kindly permitting him to carry on the research work.

References

1. Otsu, N.: A threshold selection method from gray-level histogram. *IEEE Trans. on System, Man, and Cybernetics* 9, 62–69 (1979)
2. Sauvola, J., Pietkainen, M.: Adaptive document image binarization. *Pattern Recognition*, 225–236 (2000)
3. Gatos, B., Pratikakis, I., Perantonis, S.J.: Adaptive degraded document image binarization. *Pattern Recogniton*, 317–327 (2006)
4. Valverde, J.S., Grigat, R.R.: Optimum Binarization of Technical Document Images. In: *Proceedings of IEEE International Conference on Image Processing*, vol. 3, pp. 985–988 (2000)
5. Zhang, Z., Tan, C.L.: Recovery of Distorted Document Images from Bound Volumes. In: *ICDAR*, p. 429 (2001)
6. Milewski, R., Govindaraju, V.: Binarization and cleanup of handwritten text from carbon copy medical form images. *Pattern Recognition* 41, 1308–1315 (2008)

7. Nandy (Pal), M., Saha, S.: An Analytical Study of Different Document Image Binarization Methods. In: Proceedings of IEEE National Conference on Computing and Communication Systems (COCOSYS 2009), UIT, Burdwan, January 02-04, pp. 71–76 (2009)
8. Saha, S., Basu, S., Nasipuri, M., Basu, D.K.: A Novel Scheme for Binarization of Vehicle Images Using Hierarchical Histogram Equalization Technique. In: Proceedings of 1st International Conference on Computer, Communication, Control and Information Technology (C3IT 2009), Academy of Technology, Adisaptagram, February 06-07, pp. 270–275 (2009) arXiv:1003.6059
9. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Pearson Education Asia (2002)

License Plate Localization Using Vertical Edge Map and Hough Transform Based Technique

Satadal Saha¹, Subhadip Basu², and Mita Nasipuri²

¹ MCKV Institute of Engineering, Liluah, Howrah, India

² Jadavpur University, Kolkata, India

{satadalsaha, subhadip}@ieee.org, mnasipuri@cse.jdvu.ac.in

Abstract. Automatic License Plate Recognition (ALPR) system is the growing need by government authorities of different developing countries like India for traffic monitoring and control. The purpose of ALPR system is to track the vehicles violating the speed limit or violating the traffic signal at a road crossing. License plate localization is one of the key modules of any ALPR system. The objective of the current work is to localize the license plate of the vehicle from the vertical edge map of the image using statistical distribution of the vertical edges. Hough transform is then used to tune potential area to its actual dimension. In this work, real time vehicle images are captured from a road-side surveillance camera automatically throughout day and night in an unconstrained outdoor environment. The performance of the technique is tested and it provides 91.23% accuracy when compared with the ground truth data.

1 Introduction

Automatic License Plate Recognition (ALPR) system has already been used in most of the developed countries during last decades or so. It has now being the growing need by government authorities of different developing countries like India for traffic monitoring and control. The purpose of any ALPR system is to track down the vehicles that have violated the traffic rule in the form of violating the speed limit in a road or in the form of violating the traffic signal at a road crossing. ALPR system is also implemented at toll plaza, in car parking area and in security zones for automatic recognition of license number of the vehicles that have entered into the area for specific purposes. Out of the three main modules (license plate localization, character segmentation and character recognition) of any ALPR system, the first one i.e. license plate localization is the most important and difficult task of any ALPR system.

Various methods and techniques have already been developed during last couple of decades for the purpose for efficient localization of license plate regions from vehicular images. In general, most of the works on ALPR systems use the edge property as features for localizing standardized license plate regions. In Greece, the license plate uses shining plate. The bright white background is used as a characteristic for license plate in [1]. A work on localization of Iranian license plate is done in [2]. In [3], W. Jia used mean shift algorithm for localization of license plate giving satisfactory result for license plates having color different from the body color.

Some of these works [2][3] capture the image of a vehicle carefully placed in front of a camera occupying the full view of it and taking a clear image of the license plate.

But in an unconstrained outdoor environment there may be huge variations in lighting conditions/ wind speed/ pollution levels/ motion etc. that makes localization of true license plate regions even more difficult. An exhaustive study of plate recognition is done in [4] for different European countries. An FPGA based license plate recognition system is reported in [5] using real time video processing. It uses low memory and is relatively faster than computer based system. A morphology based method is proposed in [6] giving high accuracy of localization with slightly lower recognition accuracy. An efficient fuzzy based system is reported in a work [7] for localization of license plate and Kohonen's self organizing neural model is used for recognition of characters. A two stage hybrid recognition system combining statistical and structural features is proposed in [8]. Llorens [9] used HMM to recognize the characters providing an accuracy of 92%.

In the developed countries as well as in some of the developing countries the attributes of the license plates, e.g. the size of the plate, color of the plate, font face/ size/ color of each character, spacing between subsequent characters, the number of lines in the license plate, script etc., are strictly maintained. However, in India, the license plates of the vehicles are not yet standardized across different states, making the localization and subsequent recognition of license plates extremely difficult. This large diversity in the features set of the license plate of Indian vehicles makes its localization a challenging problem to the research community.

Unfortunately, limited works [10] have been done on localizing the license plates from Indian vehicles. RGB-HSI color model is used in [11] for localization of license plate. In [12], RGB-HSI color values are used as features to train an Artificial Neural Network (ANN) and subsequently used it to localize the license plate. An efficient vertical edge detection based method is discussed in [13] to localize the license plate. In our earlier work [14], Hough transform is used as a generalized text segmentation technique to localize license plate of a vehicle.

Keeping view of the above facts, the objective of the current work is to present a robust technique for localization of license plate regions from Indian vehicular images captured from road side camera in an unconstrained outdoor environment. It is to be mentioned that the main algorithm for the localization technique used in the current work has already been reported as our earlier work [13]. Hough transform is blended here as an innovative approach to improve the efficiency of our earlier work [13].

2 Present Work

The current technique has been developed as a part of an ALPR system on demand by a state government authority of India. Surveillance cameras are installed at a road crossing and images are captured from road side cameras through out the day and night. The images are then processed using some quality improvement techniques to facilitate the localization of the license plate. Vertical edge map of each image is created using Sobel's vertical edge operator. The potential license plate regions are then localized by statistically analyzing the concentration of the vertical edges. The localized potential license plate areas are then refined or eliminated by analyzing the

Hough image. A rule based decision is then taken on the overlapping region of the vertical edge and the potential region marked in the Hough image to finally identify the true license plate region.

2.1 Image Acquisition

The image dataset for the current experiment is collected as a part of a demonstration project on automated Red Light Violation Detection System (RLVDS) for a state Government authority of a major metro city in India. Using cantilever system three surveillance cameras were installed at an important road crossing in Kolkata at a height of around ten meters from the road surface. All the surveillance cameras were synchronized with the traffic signaling system and were focused on the stop-line such that the cameras captured the snapshots only when the traffic signal was turned RED. The complete image dataset comprises of more than 25,000 still snapshots, captured over several days/nights in an unconstrained outdoor environment with varying lighting conditions, pollution levels, wind turbulences and vibrations of the camera. 24-bit color bitmap images were captured with a rate of 25 fps and resolution of 704x576 pixels. Not all these still snapshots contain vehicle images with a clear view of license plate regions. For the current work, a dataset consisting of 1500 images has been developed that contain complete license plate regions appearing in different orientations in the image frame, keeping in mind that the dataset contains variety of vehicles in terms of orientation of the vehicle, size of the vehicle, number of lines in the license plate, variation in lighting condition etc.

2.2 Pre-processing

As described in previous section, true color still snapshots of resolution 704x576 pixels were captured through multiple surveillance cameras over day and night with embedded noise and huge variations in image quality. Following preprocessing techniques are implemented in the current work to address the issues mentioned above.

Rotation: Due to the fixation of the camera, normally the front face of it makes some angle with the frontal vertical plane of the road and thereby making the license plate of the vehicle stopped at the stop-line skewed. This angle remains fixed as long as the camera remains fixed at the position of its mount. To localize the license plate the image should be rotated such that the license plate becomes horizontal in the image plane [15]. The images for a specific camera are rotated through the fixed angle corresponding to the particular camera using the formula given in equation (1).

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \quad (1)$$

Grey scale conversion: For each pixel, 24-bit color value is converted to 8-bit grey value using the formula [16] written in equation (2).

$$gray(i, j) = 0.59 * R(i, j) + 0.30 * G(i, j) + 0.11 * B(i, j) \quad (2)$$

where, (i, j) indicates the position of a pixel in the image and $gray(i, j) \in (0, 255)$.

Median filtering: As the proposed technique is based on detecting the edge of the image, salt-and-pepper noise generated randomly needs to be removed from the image. Median filter is a non-linear filter which removes salt-and-pepper noise from the image. This filter replaces the gray value of a pixel by the median of the gray values of its neighbors. In this work, a 3×3 mask is used as median filter

2.3 Edge Map Generation

Sometimes edges provide more information than the original color or grey valued images. Edge map is the binarized edge gradient computed using some edge detection operator. The objective of the current work is to find the edges created by the characters within the license plate. Sobel’s edge operator [16] is used for detection of edge gradients. It is seen that when the characters of the license number are written horizontally, the vertical edges appearing due to the presence of the license plate characters appear in very concentrated form and they have more or less a specific height. The pattern and concentration of the vertical edges also remain in conformity with the pattern of the license number. This statistical distribution of vertical edge pattern is seen to occur within the license plate of the vehicle and no where else within the natural scene of the image. In the present work, this phenomenon is explored to find the license plate region within the image.

$$gradV(y,x) = \sqrt{\left(\sum_{m=-1}^{+1} \sum_{n=-1}^{+1} Vmask(n,m) img(y+n,x+m)/4\right)^2} \tag{3}$$

The formula for getting vertical edge gradient is written in equation (3), where, *img* is the image over which the edge detection algorithm is operated upon, *Vmask* is the Sobel’s vertical edge operator as given below in equation (4) and *gradV* is the vertical edge gradient.

$$Vmask = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \tag{4}$$

Fig. 1 shows the result of pre-processing and edge detection. Fig. 1(c), (d) show that the vertical edge map is more informative in detecting the license plate as condensed vertical edges appear at the region of the license plate due to the characters therein.

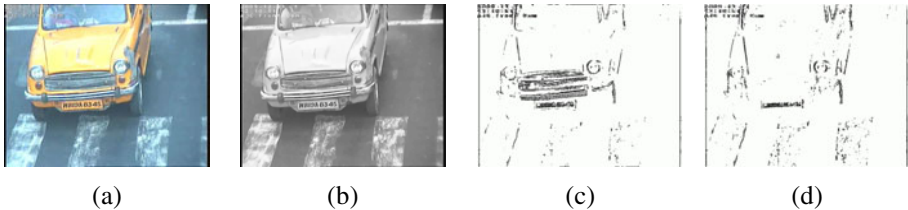


Fig. 1. (a) Original colored image, (b) Preprocessed grey image, (c) Binarized edge image, (d) Binarized vertical edge image

2.4 License Plate Localization

It may be observed from Fig. 1, that due to the presence of the characters of the license plate, the pattern of the vertical edges at the license plate region is very dense and prominent. Also, the vertical run-lengths of edge pixels within the license plate regions are almost equal to the height of the characters therein.

Using the aforesaid attributes, the overall localization algorithm may be subdivided into the following intermediate stages.

- *Identification of potential band of rows*
 - In this stage the band of rows are found where the vertical edges are continuous and reaches a particular height.
- *Primary localization of license plate regions based on statistical distribution of vertical edge pixels*
 - In this stage for each band of rows, the mean and standard deviation of the positions of the vertical edge pixels are calculated. The height of the potential region is defined as the maximum height of the vertical edge pixel for that band. The width of the region is defined as the twice the standard deviation of the position of the vertical edge pixels around the mean position.
- *Refinement of license plate regions based on prominent vertical edges*
 - Within each specified region, the first prominent vertical edge from the left side with acceptable height is set as the left boundary of the region and similarly, the first prominent vertical edge from the right side with acceptable height is set as the right boundary of the region.
- *Localization of license plate by removing the noise segments*
 - The area and the aspect ratio of each region are calculated and the regions for which the values are within some specified range are considered as the potential license plate regions.

The detailed steps have been discussed in our earlier work [13] in an algorithmic approach. The outcome of this module is potential license plate regions marked by bounding boxes, as shown in Fig. 2(b) and 3(b).

2.5 License Plate Selection

These boxes are actually selected depending on the area and aspect ratio of them. It is seen that some of the boxes though have area and aspect ratio within an acceptable range; still they do not contain any string of texts. This is formed because some insignificant and uncorrelated vertical edges form a region with acceptable area and aspect ratio but actually the vertical edge pattern is not in conformity with the pattern formed by string of text characters. The widths of the regions are based on the statistical values of mean and standard deviation of the positions of the vertical edges. This introduces extra space in the left and the right sides of the selected regions. To shrink the dimension of the selected region sidewise Hough transform [16] is used inside the regions to identify the desired regions where the vertical edges are very close to each other and select the license plate accordingly.

Hough transform is applied over each region to generate the Hough image of it, as shown in Fig. 2(d) and 3(d). For this purpose, the parameters of the Hough transform, like *deltaRo*, *deltaTheta*, *startTheta*, *endTheta*, *connectDistance* and *pixelsCount* are set such that neighboring characters become connected to each other. In the proposed work, directional Hough transform is used to mark any text line having a skew angle of 0^0 to 1^0 with respect to the horizontal axis with *deltaTheta* taken as 180^0 . The *connectDistance* and *pixelsCount* values are kept as 5 and 2 respectively. The connected component labeling (CCL) algorithm is applied over the Hough image to segment it terms of marked regions.

For each region, the Hough image and the vertical edge map are placed back to back to find the overlapping region of the vertical edge map and the segments as marked in the Hough image. If there is sufficient intersection between them and the region marked in the Hough image has acceptable area and aspect ratio then only the Hough segments are finally considered as the localized license plates.

3 Results

The experiment is run as a part of full ALPR system that was run throughout the day and night in an unconstrained road crossing. Fig. 2 and 3 show three sample cases of license plate localization with variation of lighting condition. The (b) and (c) parts of the figures show the localized license plates containing extra spaces in the left and right, as described in section 2.4. The (d) part of the figures show the vertical edge maps within the localized region and the (e) part of them show the Hough images of the localized regions. Analyzing the Hough image and the vertical edge map, the final localized license plate region is shown in (f) part of the figures.

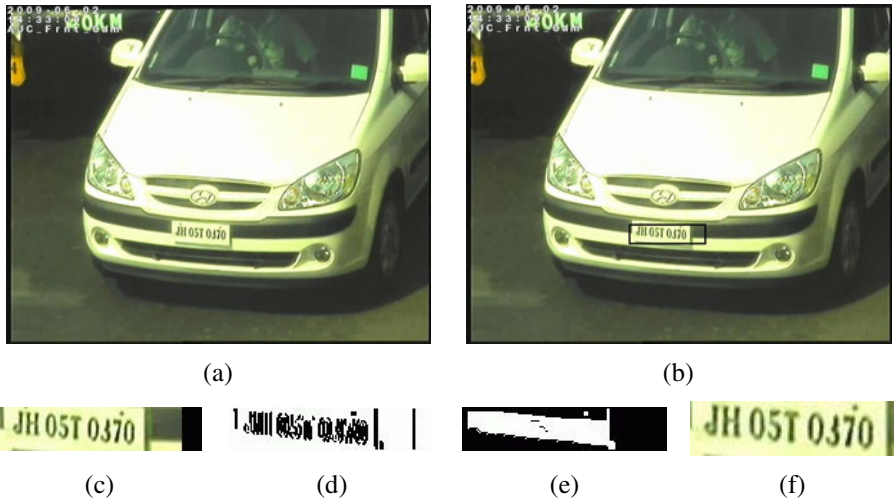


Fig. 2. Localization result for vehicle with single line license plate (a) Original image, (b) License plate indicated by bounding box, (c) Extracted license plate, (d) Vertical edge map of extracted region, (e) Hough image of the extracted region, (f) Finally selected license plate

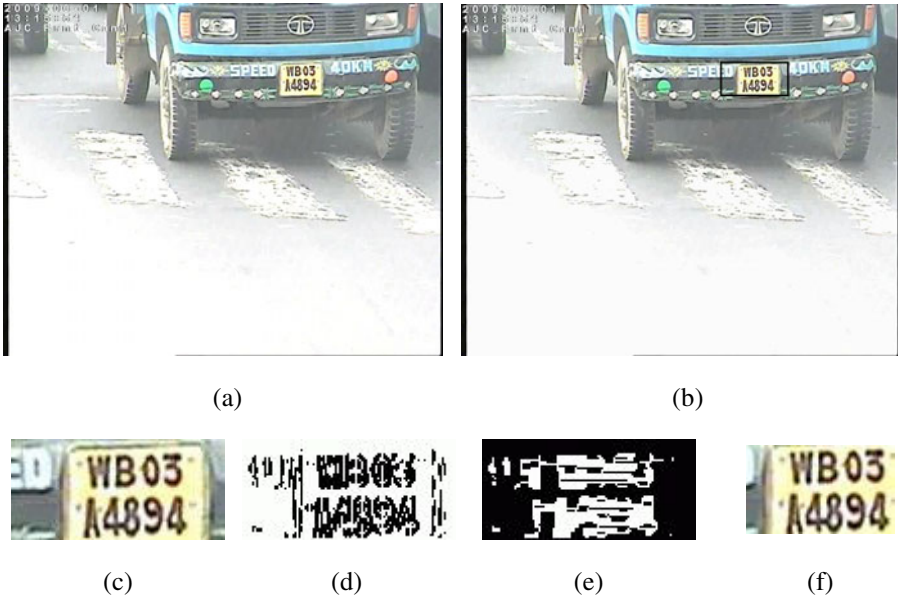


Fig. 3. Localization result for vehicle with two-line license plate (a) Original image, (b) License plate indicated by bounding box, (c) Extracted license plate, (d) Vertical edge map of extracted region, (e) Hough image of the extracted region, (f) Finally selected license plate

Comparison of the (c) and (f) parts of Fig. 2 and 3 shows that the large amount of adjacent area of the license plate can be removed from the localized region using the Hough transform. The technique is suitable to localize double line license plate also, as shown in Fig. 3.

4 Conclusion

The technique performs well in localizing the license plate of the vehicle as revealed by exhaustive unconstrained outdoor experimentation. The dataset used here contains huge variation in terms of vehicle size and type, lighting condition, license plate character font face and type, weather condition etc. The technique can identify true license plates only (*true positive*) in 91.23% cases. In 1.8% cases the true license plate is identified along with falsely localized regions (*false positive*). And in 6.97% cases no license plate is found (*false negative*). The reason behind the *false negative* cases is insufficient generation of vertical edges because of the very poor quality of the license plate. Though the basic technique used here follows our earlier work [13], more stress is given in the current work in reducing the localized license plate area and eliminating the falsely localized license plate area.

Acknowledgments. Authors are thankful to the CMATER, the SRUVM project, and the PURSE program of C.S.E. Department, Jadavpur University, for providing necessary infrastructural facilities during the progress of the work. Mr. S. Saha is thankful to the authorities of MCKV Institute of Engineering for kindly permitting him to carry on the research work.

References

1. Kawasnicka, H., Wawrzyniak, B.: License Plate Localization and Recognition in Camera Pictures. In: AI-METH 2002, Poland (November 2002)
2. Mahini, H., Kasaei, S., Dorri, F., Dorri, F.: An Efficient Features-Based License Plate Localization Method. In: Proceedings of 18th International Conference on Pattern Recognition (2006)
3. Jia, W., Zhang, H., He, X., Piccardi, M.: Mean Shift for Accurate License Plate Localization. In: Proceedings of 8th International IEEE Conference on Intelligent Transportation Systems, Vienna, Austria (September 2005)
4. Anagnostopoulos, C.N., Anagnostopoulos, I., Loumos, V., Kayafas, E.: A license plate recognition algorithm for Intelligent Transport applications. *IEEE Transaction on Intelligent Transport Systems* 7(3) (2006), <http://www.aegean.gr/culturaltec/canagnostopoulos/cv/T-ITS-05-08-0095.pdf> (2009)
5. Caner, H., Gecim, H.S., Alkar, A.Z.: Efficient Embedded Neural-Network-Based License Plate Recognition System. *IEEE Transactions on Vehicular Technology* 57(5), 2675–2683 (2008)
6. Kasaei, S.H., Kasaei, S.M., Kasaei, S.A.: New Morphology-Based Method for Robust Iranian Car Plate Detection and Recognition. *International Journal of Computer Theory and Engineering* 2(2), 1793–8201 (2010)
7. Chang, S., Chen, L., Chung, Y., Chen, S.: Automatic License Plate Recognition. *IEEE Transaction on Intelligent Transport Systems* 5(1), 42–53 (2004)
8. Pan, X., Ye, X., Zhang, S.: A Hybrid Method For Robust Car Plate Character Recognition. *Engineering Applications of Artificial Intelligence*, 963–972 (2005)
9. Llorens, D., Marzal, A., Palazon, V., Vilar, J.M.: Car License Plate Extraction and Recognition Based on Connected Components Analysis and HMM Decoding. In: Marques, J.S., Pérez de la Blanca, N., Pina, P. (eds.) *IbPRIA 2005*. LNCS, vol. 3522, pp. 571–578. Springer, Heidelberg (2005)
10. Nathan, V.S.L., Ramkumar, J., Kamakshi, P.S.: New approaches for license plate recognition system. In: *ICISIP*, pp. 149–152 (2004)
11. Saha, S., Basu, S., Nasipuri, M., Basu, D.K.: An Offline Technique for Localization of License Plates for Indian Commercial Vehicles. In: Proceedings of IEEE National Conference on Computing and Communication Systems (COCOSYS 2009), UIT, Burdwan, January 02-04, pp. 206–211 (2009), arXiv:1003.1072
12. Saha, S., Basu, S., Nasipuri, M., Basu, D.K.: Localization of License Plates from Surveillance Camera Images: A Color Feature Based ANN Approach. *International Journal of Computer Applications* 1(23), 27–31 (2010) ISSN: 0975 – 8887
13. Saha, S., Basu, S., Nasipuri, M., Basu, D.K.: License Plate localization from vehicle images: An edge based multi-stage approach. *International Journal of Recent Trends Engineering* 1(1), 284–288 (2009) ISSN 1797-9617
14. Saha, S., Basu, S., Nasipuri, M., Basu, D.K.: A Hough Transform based Technique for Text Segmentation. *Journal of Computing* 2(2), 134–141 (2010) ISSN: 2151-9617, arXiv: 1002.4048
15. Saha, S., Basu, S., Nasipuri, M., Basu, D.K.: Development of an Automated Red Light Violation Detection System (RLVDS) for Indian vehicles. In: Proceedings of IEEE National Conference on Computing and Communication Systems (COCOSYS-2009), UIT, Burdwan, January 02-04, pp. 59–64 (2009), arXiv:1003.6052
16. Gonzalez, R.C., Woods, R.E.: *Digital Image Processing*. Pearson Education Asia (2002)

Hierarchical Cluster Based Query-Driven Routing Protocol for Wireless Sensor Networks

Soumyabrata Saha¹ and Rituparna Chaki²

¹ Department of Information Technology,
JIS College of Engineering, West Bengal, India
som.brata@gmail.com

² Department of Computer Science & Engineering,
West Bengal University of Technology, West Bengal, India
rituchaki@gmail.com

Abstract. Wireless Sensor Networks require robust wireless communication that are energy efficient and provide low latency. Various routing schemes have been presented in order to improve the life time of these wireless sensor networks and to overcome the energy constraint of sensor nodes. One of the effective schemes is based on clustering of sensor nodes and as well as to improve the network life time, decreases the communication latency and to reduce the energy consumption of whole Wireless Sensor Networks as much as possible. In this paper we include a brief survey of the state-of-the-art of various existing cluster based routing algorithms for sensor networks and present a new clustering algorithm named HCBQRP: Hierarchical Cluster Based Query-Driven Routing Protocol for Wireless Sensor Networks. The main goal of this routing protocol is to evenly distribute the energy load among the entire sensor nodes in the network so that there are no overly utilized sensor nodes that will run out of energy before the others.

Keywords: Wireless sensor networks, wireless communication, query-driven, routing, hierarchical, sensor nodes, clustering.

1 Introduction

Wireless sensor networks are consists of large number of tiny autonomous sensor nodes with limited energy resources. These sensor nodes form together an ad hoc distributed sensing and data propagation network. Wireless sensor networks (WSN) consists of sinks and sensors. Sensor should send their collected data to a determined node called Sink. Sinks may use the collected data immediately or transmit them to users through a gateway. Nodes using routing protocol determine a path for sending data to sink. Though the sensor nodes carry limited and irreplaceable power source the protocols designed for the WSN must take the issue of energy efficiency into consideration.

Clustering in wireless sensor networks provides scalability and robustness for the network; it allows spatial reuse of the bandwidth, simpler routing decisions, and

results in decreased energy dissipation of the whole system by minimizing the number of nodes that take part in long distance communication. When the network is partitioned into clusters data transmission can be classified into two stages, i.e., intra-cluster communication and inter-cluster communication. Non-cluster-head nodes first send their data to the cluster heads (CHs) and CHs send the data to BS. Cluster heads co-operate with each other to forward their data to the base station. The goal is to select cluster heads that minimize transmission costs and energy usage. Cluster based routing has been shown to be more energy efficient and increase the network lifetime through data aggregation. Sensor network used to be designed as hierarchical clustering structure to achieve this goal.

In this paper we design and implement a real time routing protocol for wireless sensor networks. The proposed routing protocol is composed of four different phases. At first cluster formation and cluster head selection mechanism has been executed. The main motive of dynamic cluster head rotation mechanism is to evenly distribute the energy load among all the sensor nodes so that there are no overly-utilized sensor nodes that will run out of energy before the others. In the next step real time route tree formation has been formed among different cluster heads. This phase is highly efficient and it takes an important part for rest of the algorithm. In the third stage information sharing has been performed. At last route maintenance mechanism takes also an important role. The route maintenance phase can quickly and easily discover multiple alternate routes to deliver the data packets from sender to receiver.

The rest of the paper is organized as follows. A comprehensive survey of related works of different routing techniques in WSNs is presented in Section 2. In Section 3 We have design and describe a new simple reliable routing algorithm, which incurs the transitions among these four different phases and improves energy efficiency to prolong the whole network lifetime. Intensive result analysis is presented in section 4. Finally, we conclude our paper with final remarks in Section 5.

2 Related Works

In recent years cluster based architectures are one of the most suitable solutions, in order to cope with the requirements of large scale WSN. In this section we take a brief look at some of the common clustering algorithms applicable for WSN.

Heinzelman proposed a hierarchical clustering algorithm for sensor networks; named Low Energy Adaptive Clustering Hierarchy [16], [4]. Two layers architecture introduced in LEACH [16], [4]. One used for communication within the clusters and the other was between the cluster heads and sink. The main disadvantage with LEACH [16], [4] protocol is, due to the random selection of cluster head, there exists the probability that the cluster heads formed are unbalanced and may be present in one part of the network, making other portion of the network unreachable.

Threshold Sensitive Energy Efficient Protocol [14] pursues a hierarchical approach along with the use of a data centric mechanism. In TEEN [14] Cluster head uses hard threshold & soft threshold values. TEEN [14] is not suitable for periodic reports based applications.

Adaptive Threshold sensitive Energy Efficient Sensor Network Protocol [12] is an extension to TEEN [14] and aims at both capturing periodic data collections and

reacting to time critical events. The main features of the APTEEN [12] scheme is that; it combines both proactive and reactive policies. The main drawback of the scheme is the additional complexity required to implement the threshold functions and the count time.

Lindsey, Raghavendra were proposed a Power Efficient Gathering in Sensor Information Systems [13], which is a chain based routing protocol. PEGASIS [13] increases the lifetime of each node by using collaborative techniques and as a result the network lifetime will be increased. It also allows only local coordination between nodes that are close together, so that less bandwidth has been consumed in communication. PEGASIS [13] introduces excessive delay for distant node on the chain and in addition the single leader can become a bottleneck.

Boukerche, Pazzi and Araujo were proposed a Periodic Event Driven and Query Based Routing Protocol [9], [8] consists of three steps: the construction of the hop tree, i.e., the dissemination tree; the propagation of subscriptions; and the data delivery to the Sink. The basic idea of the PEQ [9], [8] is that it uses the hop level of the nodes as the main information to minimize data transmission.

The Clustering PEQ [9], [8] was designed based on the PEQ [9], [8] mechanism. CPEQ [9], [8] employs an energy aware cluster head selection mechanism in which the sensor nodes with more residual energy are selected to become cluster heads. The CPEQ [9], [8] protocol configures the dissemination tree using the PEQ's [9], [8] algorithm with a simple modification that is an additional field that contains the percentage of nodes that can become CHs.

An Energy Efficient Inter Cluster Communication based Routing Protocol for WSNs [9], [8] does not determine whether the nearest neighbor nodes are able to communicate or not. ICE [9], [8] also provides QoS by finding a path with the least cost for high priority event notification messages.

Hybrid Energy Efficient Distributed [11] clustering introduces a variable, known as cluster radius which defines the transmission power to be used for intra cluster communication. HEED [11] terminates within a constant number of iterations and achieves fairly uniform distribution of cluster heads across the network.

Li-Ming-He was proposed a Novel Real Time Routing Protocol [3] to guarantee real time communication. Through real time route tree construction, source node can discover the optimal route along which sensing data can be delivered to sink node with minimum delay. The proposed novel phase transition mechanism ensures multiple suboptimal routes are used, which prolongs the network lifetime greatly.

Energy Efficient Hierarchical Routing Protocol [1] motivates the need for data collection in the gateway area by the one which no cluster head node to attach with. By using Energy Efficient Hierarchical Routing Protocol [1] the cluster heads can preserve some energy in data forwarding and gateway nodes can ease their burden and gateway node not taking participation in cluster formation.

Wenjun Liu and Jiguo Yu were proposed Energy Efficient Clustering and Routing Scheme [2], includes three phases: distributed nodes clustering, dynamic cluster head rotation and inter cluster routing selection. Routing selection takes advantage of base station's energy and the communication overhead and network lifetime of EEER [2] are also desirable.

In the next section we are going to propose a new routing protocol and try to reduce the problems of previously discussed routing protocols.

3 Proposed New Routing Protocol

Hierarchical routing involves in the cluster formation techniques where low energy nodes are assigned the task of sensing. The main aim of hierarchical routing is to efficiently maintain the energy consumption of sensor nodes by involving them in multi-hop communication.

We have used the first order radio model [16]. In this model, a radio dissipates 50nJ/bit (EM_{elec}) to run the transmitter or receiver circuitry and 100pJ/bit/m² (EM_{amp}) for the transmitter amplifier. The energy consumption model is described as follows. When a sensor node transmits m-bit data to another node with distance d, the energy it consumes is

$$EM_{TX}(m, d) = EM_{elec} * m + EM_{amp} * m * d^2 \quad (1)$$

After receiving m bit data, the sensor node consumes the energy,

$$EM_{RX}(m) = EM_{elec} * m \quad (2)$$

Here EM_{elec} is denoted as the circuit energy cost for transmitting or receiving one bit data and EM_{amp} is the amplifier coefficient and d^2 is the energy loss model is used for representing channel attenuation.

Before designing of the newly proposed routing protocol few assumptions have been summarized. Such that; Sensor nodes are generally energy constrained; Every sensor node has unique identifier and they can directly communicate with its immediate neighbor; Wireless sensor network is usually data centric, i.e. application specific; Normally data collections by the sensor nodes have been done based on the locality; In Wireless sensor nodes every link between any two nodes is bidirectional; The transmission range of each node is same on one condition that the transmission range should cover all the neighbors in the network. In our proposed protocol, WSN perceived as a network partitioned into clusters based on locality.

The Proposed Routing Protocol is composed of four different phases.

- I. Cluster Formation and Cluster Head Selection Process.
- II. Real Time Route Tree Formation Among Different Cluster Heads.
- III. Information Sharing.
- IV. Route Maintenance Mechanism.

I. Cluster Formation and Cluster Head Selection Process

The sensor nodes of the same location are usually collect almost same redundant data. To reduce data redundancy cluster heads apply data aggregation mechanism between collected data and put into a single length fixed packet. This mechanism is suitable for saving the energy of the sensor nodes without transferring the redundant data.

For applying the above discussed mechanism, in our proposed protocol we are using the clustering technique. Here local nodes are involved to form the cluster and those nodes have the highest energy level, will elect as a cluster head. This protocol implements randomized rotation of CHs to evenly distribute the energy load among the different sensor nodes in the network. It helps prevent draining the energy of the sensor node and as well as this dynamic clustering increases the network lifetime. When the energy of a cluster head drops below the threshold, it replaced by a with a

new cluster head. We assume that all sensors are identical and produce data at the same rate. To calculate EM_{TH} , we use these following equations:

$$m = n * p \tag{3}$$

$$EM_{CH} = EM_{elec} * m + EM_{amp} * m * d_{CH}^2 \tag{4}$$

Where, m as the total length of the received message in the cluster-head assuming there are n member nodes in the cluster and each message is p-bit long and d_{CH}^2 is the distance between the cluster-heads. EM_{CH} is the energy consumed as the cluster-head transmits the aggregated data to other cluster heads. The energy consumed by a cluster head which is obtained by Equation (4). In Equation (5) average energy consumption per cluster-head has been calculated,

$$EM_{TH} = 1/c \sum_{i=1}^c EM_{CH} \tag{5}$$

Where, c is the no of clusters. Since EM_{TH} changes over time, the threshold is calculated in every data collection and transmission phase.

II. Real Time Route Tree Formation Among Different Cluster Heads

We are trying to form a route among different cluster heads rather than different sensor nodes. By sharing information only among the cluster heads, they can acquire all the information which is sensed by the sensor nodes.

Any sensor node in the wireless sensor network with multiple interests passes those to its cluster head. Then that cluster head becomes sink node and it enters the real time route tree as a root of the tree and assigns itself to level 0. Then the sink node broadcast a **RT_FRM_MSG {Clsr_Hd_Id, Msg_Id, Level, Int₁, Int₂... Int_n, TTL}** to its neighbor cluster heads.

Each cluster head maintains a **MSG_HS_TBL**. Once any cluster head receives **RT_FRM_MSG**, it copies all the information to its **MSG_HS_TBL {Clsr_Hd_Id, Msg_Id, Level, Int₁, Int₂,..., Int_n, TTL, Tm_Stmp}** and records the time of the message arrival in the time stamp field and assigns itself a level which is one greater than the level of the message from where it has been received. In this way, this cluster head becomes the child of the previous cluster head. This above process is applicable for all the remaining cluster head in the network.

After entering a route formation tree if any node again receives same **RT_FRM_MSG** then store the message details in the **MANTAN_TBL** otherwise copy the details to the **MSG_HS_TBL**, i.e. the real time route has been discovered between different cluster heads.

From the fig1, we are assuming that a sensor node of Cluster-A having the multiple interests and it passes its interests to its cluster head (CH_A). Then CH_A becomes the sink node and enters in the real time route formation tree and becomes the root of the tree and assign its level to 0.

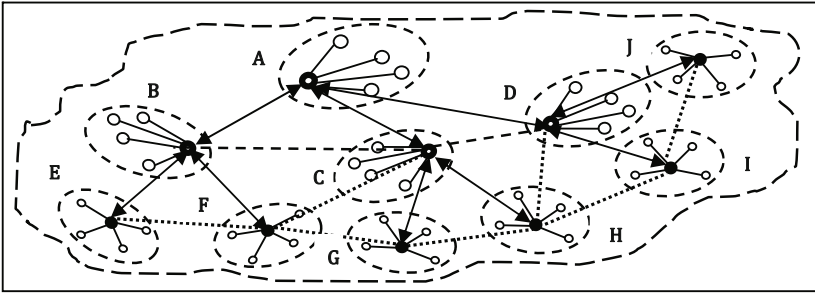


Fig. 1. Real time route formation process

In the next steps this sink node CH_A broadcasts a RT_FRM_MSG to its neighbor Clusters, i.e. Cluster-B, Cluster-C & Cluster-D. After receiving the RT_FRM_MSG they have copied all these information to their MSG_HS_TBL & as well as they have updated their level, one greater than the level of the message. Now the levels of these clusters heads are 1. By using this technique all these cluster heads, i.e. CH_B , CH_C , CH_D are become the child of the CH_A . This discussed process is applicable for the entire network.

III. Information Sharing

In our proposed protocol we have to address two different types of Information Sharing mechanism.

A. Inter Cluster Information Sharing

In our proposed protocol we have three types of inter cluster communications.

- a. Cluster members are wanted to share information with their own cluster head.
- b. Multiple interests are generated by the cluster members and those are passes to their own cluster head.
- c. After getting some desired information for satisfying the interest of the cluster member, cluster head forwards that information to the cluster member.

Time-driven fashion is applicable for case a. Both case b and case c follow event-base passion.

B. Intra Cluster Information Sharing

In this case, information sharing has been occurred among different cluster head. When cluster head sensed any event, it searches for corresponding interest in its MSG_HS_TBL {**Sens_Data**, **Msg_Id**, **Snk_Nd_Id**, **Src_Nd_Id**, **TTL**}. If any interest is present, then the cluster head becomes source node and begin to send sensing data to the sink node, i.e. to the root cluster head which belong in the level 0.

From fig2, we are observing that the CH_A passes its multiple interests to its neighbor CH_B , CH_C and CH_D . These cluster heads also forwards RT_FRM_MSG to all the neighbor clusters, updating level to 1. Now CH_B , CH_C & CH_D have copied all the duplicate RT_FRM_MSG to the $MANTAN_TBL$

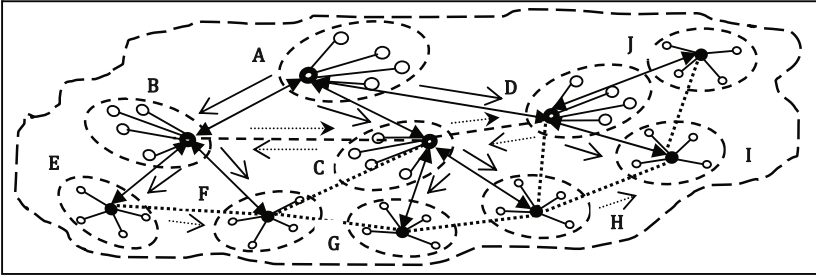


Fig. 2. Communication within real time route tree

In the level two clusters i.e. CH_H , CH_I & CH_J are received this message from level 1 cluster head, CH_D and copy that particular message to its MSG_HS_TBL . As well as these level 2 cluster heads are forwarded the RT_FRM_MSG to each other. All these three cluster heads copy the duplicate RT_FRM_MSG to the $MANTAN_TBL$.

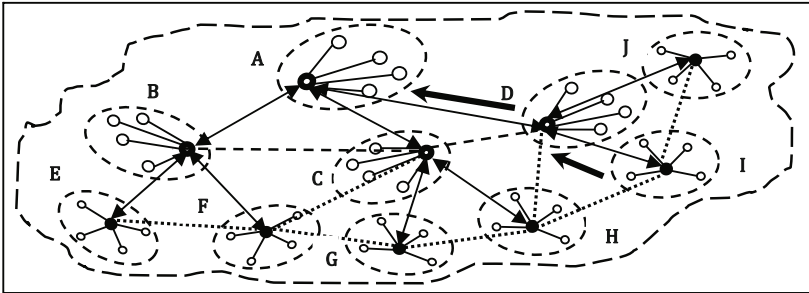


Fig. 3. Message delivery process within real time route tree

We are considering that is the node CH_I who is able to solve the interest of the sink cluster head CH_A . The interests are reached from sink CH_A in level 0 to source CH_I in level 2 via this specified path, i.e., $CH_A \rightarrow CH_D \rightarrow CH_I$

After building up the route from sink node CH_A to source node CH_I , the source node sends the solution to the sink node via the same reverse path, i.e. $CH_I \rightarrow CH_D \rightarrow CH_A$.

IV. Route Maintenance Mechanism

In the route, if any sensor node will damage for a certain time period then this entire path will damage, to handle this kind of situation we have incorporated route maintenance mechanism.

After receiving any information from any sensor node, the receiving node has send ACK_MSG to the sender node within a certain time period. These $ACK_MSG\{Snk_Nd_Id, Recvr_Nd_Id, Msg_Id, Ack_Id, Tm_Stmp\}$ based repair mechanisms consist of two parts. One is Failure Detection and other is selection of Alternative node.

Depends on the reception time of Ack_Msg sensor nodes can detect if its neighbor nodes are functioning properly or not. If Ack_Msg not received within a specified time, the sensor node search in the MANTAN_TBL to find out one alternative cluster head from which it has receives same RT_FRM_MSG. If no such information is found in MANTAN_TBL the entire process is terminated, otherwise it will find out the alternative route.

By using the above discussed algorithm any sensor node can communicate with any another nodes through the cluster head within that network.

Table 1. Data Dictionary

Data Value	Description
Clsr_Hd_Id	Cluster Head Id
CH _x {x=A to J}	Cluster Head
Msg_Id	Message Id
Level	Level of the Cluster
Int _i	Interest Number
TTL	Time to Leave
Tm_Stmp	Time Stamp
Sens_Data	Sensing Data
Snk_Nd_Id	Sink Node Id
Src_Nd_Id	Source Node Id
Snd_Nd_Id	Sender Node Id
Recvr_Nd_Id	Receiver Node Id
Ack_Id	Acknowledgement Id

/* algorithm for cluster head selection */

√ **Clsr_Hd_Slect ()**

Step1: Set cluster head \leftarrow n1, Energy level of cluster head \leftarrow E_{CH}

Step2: Broadcast a message with E_{CH}

Step3: IF (E_{CH} < node n energy level)

 StepI : node n sends a message with it's energy level to n1

 StepII: node n elected as cluster head

ELSE node n1 remain cluster head.

/* algorithm for route formation */

√ **Rt_Frm_Inf_Shr ()**

Step1: Set sink_node \leftarrow cluster head of cluster node with multiple interest

Step2: Set Level_{sink_node} \leftarrow 0

Step3: sink_node sends RT_FRM_MSG

Step4: IF the receiver node not entered in that route & can't satisfy interest

 StepI : update MSG_HS_TBL with RT_FRM_MSG

 StepII: add message arrival time

 Step III: update RT_FRM_MSG with

 Step a: level = level+1

 Step b: Clsr_Hd_Id \leftarrow it's own id.

```

StepIV: forward RT_FRM_MSG to it's neighbor
ELSE IF the receiver node can satisfy interest
    THEN this node become the source node and sends a reply to that
        node from which it has receives the message
    ELSE IF the receiver node is already entered in that route
        IF the message is same as previous one
            Update MANTAN_TBL with RT_FRM_MSG
        ELSE
            Update MSG_HS_TBL with RT_FRM_MSG

```

/* algorithm for route maintenance */

√ **Rt_Mnt ()**

Step1: Search in MANTAN_TBL

Step2: IF any alternative path, then update route

ELSE terminate the process.

4 Result Analysis

The simulation model consists of a network model that has a number of mobile wireless node models, which represents the entire network to be simulated.

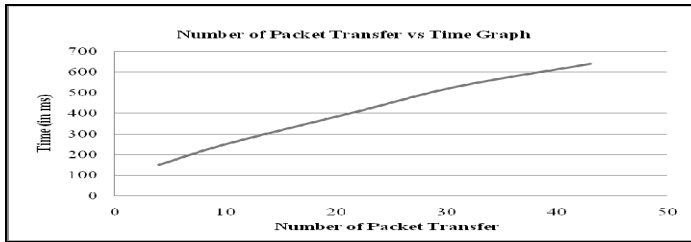


Fig. 4. Number of packet transfer vs time graph

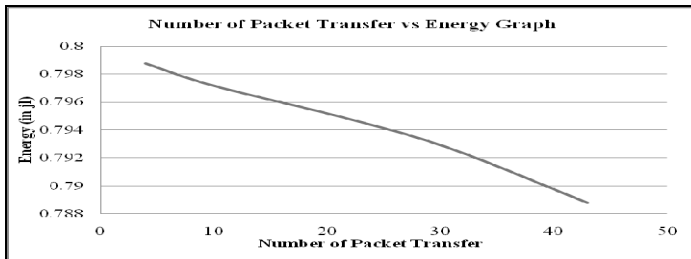


Fig. 5. Number of packet transfer vs energy graph

The graph in fig 4, shows the time required for route discovery process from source to sink. From that it is clearly seen that when more number of packets are transferred among different nodes then the required time is also increased. In fig5, we have depicted a relation between number of packet transfer and energy. Here it has been observed that when more number of packets are transferred corresponding energy are dissipated.

5 Conclusions

In this paper, we have summarized the generic characteristics of some well known hierarchical cluster based routing protocols for WSNs and present a new routing scheme named Hierarchical Cluster Based Query-Driven Routing Protocol for wireless sensor networks to achieve real-time communication and high energy efficiency. The cluster head of each cluster acts as a local coordinator for its cluster, performing inter-cluster routing, data forwarding and has to undertake heavier tasks so that it might be the key point of the network. In sensor network most of the sensor nodes share information among all other sensor nodes within their communication range. In our proposed routing scheme only cluster head can take part to share information with their neighboring cluster heads. As a result, energy minimization has been performed. This HCBQRP protocol is also follow a query driven mechanism, i.e. when any sensor node generate any query then only information sharing has been performed through different cluster heads. Some result analysis are also incorporated to show in which way the proposed algorithm work to achieve the energy efficiency, scalability, information sharing and route maintenance mechanism.

References

1. Li, H., Shi, J., Yang, Q., Zhang, D.: 'An Energy-Efficient Hierarchical Routing Protocol for Long Range Transmission in Wireless Sensor Networks. In: IEEE 2nd International Conference on Education and Computer (2010)
2. Liu, W., Yu, J.: Energy efficient clustering and routing scheme for Wireless Sensor Networks. In: IEEE Conferences (2009)
3. He, L.-M.: A Novel Real Time Routing Protocol for Wireless Sensor Networks. In: 10th IEEE ACIS International Conference on Software Engineering, Artificial Intelligences, Networking and Parallel/Distributed Computing (2009)
4. Israr, N., Awan, I.U.: Multilayer Cluster Based Energy Efficient Routing Protocol for Wireless Sensor Networks. International Journal of Distributed Sensor Networks 4, 176–193 (2008)
5. Martirosyan, A., Boukerche, A., Pazzi, R.W.N.: A Taxonomy of Cluster-based Routing Protocols for Wireless Sensor Networks. In: IEEE, The International Symposium on Parallel Architectures, Algorithms, and Networks (2008)
6. Boukerche, A., Martirosyan, A.: An Energy-Aware and Fault Tolerant Inter-cluster Communication based Protocol for WSN. In: Globecom, Washington D.C (November 2007)

7. Ewa Hansen, J.N., Nolin, M., Björkman, M.: Efficient Cluster Formation for Sensor Networks, Mälardalen Real-Time Research Centre, Mälardale University, pp. 1404–3041 (2006) ISSN: 1404–3041 ISRN MDH-MRTC-199/2006–1-SE
8. Boukerche, A., Pazzi, R.W., Araujo, R.B.: Fault-tolerant wireless sensor network routing protocols for the supervision of context-aware physical environments. *Journal of Parallel and Distributed Computing, Algorithms for Wireless and Ad-Hoc Networks* 66(4), 586–599 (2006)
9. Muraganathan, D.C.F.M.S.D., Bhasin, R.I., Fapojuwo, A.O.: A centralized energy efficient routing protocol for wsns. *IEEE Communication Magazine*, 8–13 (2005)
10. Kim, K.-T., Youn, H.Y.: Energy-Driven Adaptive Clustering Hierarchy (EDACH) for Wireless Sensor Networks. In: Enokido, T., Yan, L., Xiao, B., Kim, D.Y., Dai, Y.-S., Yang, L.T. (eds.) *EUC-WS 2005*. LNCS, vol. 3823, pp. 1098–1107. Springer, Heidelberg (2005)
11. Younis, O., Fahmy, S.: HEED: A hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks. *IEEE Transactions on Mobile Computing* 3(4), 660–669 (2004)
12. Agrawal, A.M.A.D.P.: ATEEN: A Hybrid protocol for efficient routing and comprehensive information retrieval in wireless sensor networks. In: *2nd International Workshop on Parallel and Distributed Computing, in Wireless Networks and Mobile Computing* (2002)
13. Lindsey, S., Raghavendra, C.: PEGASIS: Power-Efficient Gathering in Sensor Information Systems. In: *IEEE Aerospace Conference Proceedings, March 9-16, vol. 3*, pp. 1125–1130 (2002)
14. Agrawal, A.M.A.D.P.: TEEN: A Protocol for Enhanced Efficiency in Wireless Sensor Networks. In: *Proceedings of 1st International Workshop on Parallel and Distributed Computing, in Wireless Networks and Mobile Computing* (2001)
15. Savvides, A., Han, C.-C., Srivastava, M.: Dynamic energy-grained localization in Ad-Hoc networks of sensors. In: *Seventh ACM Annual International Conference on Mobile Computing and Networking (MobiCom)*, pp. 166–179 (July 2001)
16. Heinzelman, A.C.A.H.B.W.: Energy-Efficient Communication Protocol for Wireless Micro sensor Networks. In: *Proc. 33rd Hawaii Int'l. Conf. Sys. Sci.* (2000)

Application of Dynamic Clustering Using ADE to Transportation Planning

Akundi Sai Hanuman¹, Sesham Anand², A. Vinaya Babu³, and A. Govardhan³

¹ GRIET, Hyderabad

a_saihanuman@hotmail.com

² MVSREC, Hyderabad

³ JNTU, Hyderabad

Abstract. In recent past, tremendous work has been done to find optimal number of clusters at run time for partitional clustering algorithms. Various Evolutionary Computation techniques have been used by researchers to evolve most appropriate number of clusters for different clustering problems. In this paper, we attempt to apply a new variant of adaptive differential evolution technique on a real world data set to find optimal number of clusters at runtime. The DCADE algorithm has been applied on Home Interview Survey (HIS) data related to a Transportation Project. Later clusters are formed and analyzed which are in accordance with the domain expert.

Keywords: Partitional Clustering, Adaptive Differential Evolution, Dynamic Clustering, Home Interview Survey.

1 Introduction

Evolutionary computation techniques have been extensively used in past to solve clustering problems of complex datasets. But not much impressive works have not been done to determine optimal number of clusters dynamically while solving clustering problem. The optimal number of cluster value is remaining to be a big challenge for partitional clustering algorithms like classical K-means even till date. Most of the existing clustering techniques, based on evolutionary algorithms, accept the number of classes K as an input instead of determining the same on the run. However, practically the appropriate number of groups in a new dataset may be unknown or impossible to determine even approximately.

It is challenging to find an optimal number of clusters in a large dataset. This problem has been investigated by several researches [1-2] but the outcome is still unsatisfactory [3]. Lee and Antonsson [4] used an Evolutionary Strategy (ES) [5] based method to dynamically cluster a dataset. In their approach, a variable-length individual to search for both centroids and optimal number of clusters is implemented. In [6] Sarkar et al have used Evolutionary Programming (EP) to classify dynamically the dataset. They have optimized two fitness functions simultaneously: one gives the optimal number of clusters, whereas the other leads to a proper identification of each cluster's centroid. Bandopadhyay et al. [7] devised a variable string-length genetic algorithm (VGA) to tackle the dynamic clustering problem using a single fitness function. Very recently, Omran et al. came up with an automatic hard clustering

scheme [8]. The algorithm starts by partitioning the dataset into a relatively large number of clusters to reduce the effect of the initialization. Using binary PSO [9], an optimal number of clusters is selected. Finally, the centroids of the chosen clusters are refined through the K-means algorithm. The authors applied the algorithm for segmentation of natural, synthetic and multi-spectral images. Das et al. [11] proposed an algorithm, which determines an optimal number of clusters at run time using DE algorithm. In this work we applied Dynamic Clustering Using Adaptive Differential Evolution to a real time data set related to a transportation project to know the optimal number of clusters at runtime.

2 Dynamic Clustering Method

The vector representation of the candidate solutions is the key point in dynamic clustering problem. It is important to specify the entire dimension the dataset has along with the user specified number of clusters in the vector. A vector having a length of $K_{max} + K_{max} \times d$ represents for a dataset having n data points with d -dimensions and a user-specified maximum number of clusters K_{max} . The first K_{max} entries are positive floating-point numbers in $[0, 1]$, each of which controls whether the corresponding cluster is to be activated (i.e. to be really used for classifying the data) or not. The remaining entries are reserved for K_{max} cluster centers, each d -dimensional. For example a single vector can be shown as:

$$\vec{V}_i(t) = \left[\begin{array}{c|c|c|c|c|c|c|c} T_{i,1} & T_{i,2} & \dots & T_{i,K_{max}} & \vec{m}_{i,1} & \vec{m}_{i,2} & \dots & \vec{m}_{i,K_{max}} \end{array} \right]$$

Activation Threshold
Cluster Centroids

The j th cluster center in the i th chromosome is active or selected for partitioning the associated data set if $T_{i,j} > 0.5$. On the other hand, if $T_{i,j} < 0.5$, the particular j th cluster is inactive in the i th chromosome. Thus, the $T_{i,j}$'s behave like control genes (we call them activation thresholds) in the chromosome governing the selection of the active cluster centers. The rule for selecting the actual number of clusters specified by one chromosome is

$$\begin{aligned} &\text{IF } T_{i,j} > 0.5, \text{ THEN the } j\text{th cluster center} \\ &\quad m_{i,j} \text{ cluster centroid is ACTIVE} \\ &\text{ELSE } m_{i,j} \text{ cluster centroid is INACTIVE} \end{aligned} \tag{1}$$

When a new offspring chromosome is created as per the basics of PSO or GA or DE algorithm at first, the T values are used to select [using above specified rule] the active cluster centroids. If due to mutation some threshold $T_{i,j}$ in an offspring exceeds 1 or becomes negative, it is forcefully fixed to 1 or 0, respectively. However, if it is found that no flag could be set to 1 in a chromosome (all activation thresholds are smaller than 0.5), we randomly select two thresholds and reinitialize them to a random value between 0.5 and 1.0. Thus, the minimum number of possible clusters is 2.

The quality of a partition can be judged by an appropriate cluster validity index. Cluster validity indices correspond to the statistical-mathematical functions used to evaluate the results of a clustering algorithm on a quantitative basis. Generally, a

cluster validity index serves two purposes. First, it can be used to determine the number of clusters, and secondly, it finds out the corresponding best partition. One traditional approach for determining the optimum number of classes is to run the algorithm repeatedly with different number of classes as input and then to select the partitioning of the data resulting in the best validity measure [10][11]. Ideally, a validity index should take care of the following aspects of the partitioning:

1) **Cohesion:** Patterns in one cluster should be as similar to each other as possible. The fitness variance of the patterns in a cluster is an indication of the cluster's cohesion or compactness.

2) **Separation:** Clusters should be well separated. The distance among the cluster centers (may be their Euclidean distance) gives an indication of cluster separation.

In the present work we have based our fitness function on the Xie-Benni index. This index, due to Xie and Beni [12], is given by

$$XB_m = \frac{\sum_{i=1}^c \sum_{j=1}^n u_{ij}^2 \|\bar{X}_j - \bar{V}_i\|^2}{n \times \min_{i \neq j} \|\bar{V}_i - \bar{V}_j\|^2} \quad (2)$$

Using XB_m the optimal number of clusters can be obtained by minimizing the index value. The fitness function may thus be written as

$$f = \frac{1}{XB_i(c) + eps} \quad (3)$$

where XB_i is the Xie-Benni index of the i -th particle and eps is a very small constant (we used 0.0002). So maximization of this function means minimization of the XB index.

The pseudocode for the Dynamic clustering algorithm is given here.

Step 1) Initialize each chromosome to contain K number of randomly selected cluster centers and K (randomly chosen) activation thresholds in $[0, 1]$.

Step 2) Find out the active cluster centers in each chromosome with the help of the rule described in (1).

Step 3) For $t = 1$ to t_{max} do

- a) For each data vector X_p calculate its distance metric $d(X_p, m_{ij})$ from all active cluster centers of the i th chromosome V_i .
- b) Assign X_p to that particular cluster center $m_{i,j}$, where

$$d(X_p, m_{i,j}) = \min_{b \in \{1, 2, \dots, K\}} \{d(X_p, m_{ib})\}.$$

- c) Check if the number of data points that belong to any cluster center $m_{i,j}$ is less than 2. If so, update the cluster centers of the chromosome using the concept of average described earlier.
- d) Change the population members according to the evolutionary algorithm outlined in each case (i.e. in this paper we used ADE). Use the fitness of the chromosomes to guide the evolution of the population.

Step 4) Report as the final solution the cluster centers and the partition obtained by the globally best chromosome (one yielding the highest value of the fitness function) at time $t = t_{max}$.

2.1 Adaptive Differential Evolution

DE is a population-based search strategy which involves floating-point encoding for continuous optimization. Like any other evolutionary technique it also goes through a simple cycle of stages. Generation of populations (i.e. candidate solutions), fitness evaluations, and replacement of current population with fitter individuals and repetition of above stages till the stopping criteria is achieved or desired results are obtained are the common stages in DE approach. The basic algorithm with terminologies of DE is explained below.

Let S be the search space of the problem under consideration having candidate solutions with n-dimension belonging into a real-world R. The DE evolves a population of NP n-dimensional individual vectors, i.e. solution candidates, $X_i = (x_{i1}, \dots, x_{in}) \in S, i = 1, \dots, NP$, from one generation to the next. For each feature (also known as attribute or dimension) of the problem it may have a certain range within which it values lie, the candidate vectors in the generation should be initialized within that bound as much as possible by uniformly randomizing individuals. For example

If

$$X_{min} = [x_{1,min}, x_{2,min}, \dots, x_{n,min}] \text{ and } X_{max} = [x_{1,max}, x_{2,max}, \dots, x_{n,max}]$$

then we may initialize the jth component of ith vector as

$$x_{j,i,0} = x_{j,min} + \text{rand}_{i,j}[0,1] \cdot (x_{j,max} - x_{j,min})$$

where $\text{rand}_{i,j}[0,1]$ is a uniformly distributed random number lying between 0 and 1.

At each generation G, DE applies mutation and crossover operations to produce a trial vector $U_{i,G}$ for each individual vector $X_{i,G}$, also known as target vector, in the current population. The process of mutation and crossover operations are described below.

[1] Mutation Operation

The mutation means sudden change in the gene characteristics of a chromosome. In the DE studies mutation means a change or perturbation of the elements of the vector. For each target vector $X_{i,G}$, an associated mutated vector $V_{i,G} = \{v_{1i,G}, v_{2i,G}, \dots, v_{ni,G}\}$ can be generated by using one of the 5 strategies described in [4]. In this work we used Scheme1.

Scheme 1: DE/rand/1

In this scheme, to create a donor vector $V_{i,G}$ for each ith member, three other parameter vectors (say $V_{r1,G}, V_{r2,G}, V_{r3,G}$) are chosen randomly from the current population where $i \neq r_1 \neq r_2 \neq r_3$. A scalar number F is taken which is known as mutation scale factor. This number scales the difference of any two of the three vectors and the resultant is added to the third one. For the ith donor vector, this process can be given as

$$V_{i,G} = X_{r1,G} + F \cdot (X_{r2,G} - X_{r3,G})$$

[2] Crossover Operation

The crossover operation is used to diversify the population after formation of donor vector. Here, the basic idea is to exchange the components between the target vector and donor vector to form a trial vector $U_{i,G} = \{u_{1i,G}, u_{2i,G}, \dots, u_{ni,G}\}$. There exists two kind of crossover in DE family of algorithms known as *exponential* (or two-point

modulo) and *binomial* (or uniform)[13]. In our work we have used binomial crossover scheme. In this scheme the trial vector is formed by following approach.

$$u_{j,i,G} = \begin{cases} v_{j,i,G} & \text{if } (rand_j \leq CR) \text{ or } j=j_{rand} \\ x_{j,i,G} & \text{otherwise} \end{cases}$$

Where $j=1,2,\dots,n$ and CR is a user-defined crossover constant in the range [0,1] and j_{rand} is a randomly chosen integer in the range [1, n] to ensure that the trial vector will differ from its corresponding target vector by at least one element.

[3] Selection Operation

This step is performed to keep the population size constant over subsequent generations. It means the total number of vectors In generation G and next generation G+1 remains constant, however the fitter vectors are only selected into the generation G+1. To compute the fitter individuals the fitness evaluations are done which is dependent on the problem under investigation. For example if a numerical function optimization is the problem and the objective is to find minimization then the individual vector having lower function evaluation value is preferred to the one having higher function evaluation values. In our clustering work, the vector giving minimum intra cluster distance is preferred to the one giving higher distance values. The selection operation is given below.

$$\begin{aligned} X_{i,G+1} &= \{ U_{i,G} \text{ if } f(U_{i,G}) \leq f(X_{i,G}) \\ &\text{Else} \\ X_{i,G+1} &= X_{i,G} \end{aligned}$$

In the process of selection the population either gets better (with respect to the minimization of the objective function) or remains the same in fitness status, but never deteriorates. It is clearly evident from the selection operation that even if the population is not improving over next generation, it moves over flat fitness landscape with generations by the replacement of trial vector by target vector

2.2 Control Parameters of ADE

There are three control parameters in basic DE algorithm. They are mutation scale factor F , crossover constant CR , and population size NP . These parameters are largely responsible to improve the performance of DE. A good amount of research work has gone into the setting of these parameters to obtain better performance for different types of problems. Zaharie in [14] proposed a new adaption strategy by controlling the population diversity with multi population implementation. In [14] it is observed that if the value of F is sufficiently small, the population can converge even in the absence of selection pressure. Omran *et al.* in [15] suggested a variant DE algorithm called SDE wherein the F is self-adapted and CR is generated from a normal distribution $N(0.5,0.15)$. They claimed that SDE performs better than DE over four benchmark functions [15]. In this work we have used yet another variation of adaptive mutation scale factor strategy keeping crossover constants and population size fixed to reasonable values as demonstrated in [10]. In this approach the mutation scale factor F is adapted on the variance of the population fitness of all vectors in a generation. The central idea lies in the fact that if the variance of fitness values of all vectors in a population varies to large value the search process becomes too much

random and convergence may take more time and on the other hand if the variance value is too small it will lead to premature convergence. Hence, if the mutation scale factor F is adapted based on the variance values of the population the step wise convergence can be obtained achieving the optimal solution. The mutation scale factor F is updated by finding the variance of the population fitness as

$$\left[\sigma^2 = \sum_{i=1}^M \left(\frac{f_i - f_{avg}}{f} \right)^2 \right]$$

Where f_{avg} =average fitness of the population of vectors in a given generation.

f_i =fitness of the i th vector in the population.

M =total number of vectors

$$f = \left\{ \max \left(\left| f_i - f_{avg} \right| \right) \right\}_{i=1,2,3,\dots,M}$$

Here, f is a normalizing factor to limit σ . A large value of σ will make the search random and whereas a small value of σ or $\sigma=0$, the solution tends towards a premature convergence. To alleviate this phenomenon and to obtain optimal solution, the F is updated as

$$F(k) = \lambda F(k-1) + (1-\lambda)\sigma^2$$

The forgetting factor λ is chosen as 0.9 for faster convergence.

The Pseudo-code for the ADE algorithm with binomial crossover is given in the box below

Pseudo-Code for the ADE //

Step1: Initialize the control parameters such as crossover constant CR and the population size NP.

Step2: Set the generation $G=0$, and randomly initialize individual vectors of population

Step3: WHILE the stopping criterion is not satisfied

DO

FOR $i=1$ to NP

Mutation step

Generate a donor vector corresponding to the i th target vector via the mutation scheme as described earlier in the paper. (In our work we used Scheme 1)

Crossover step

Generate trial vector for the i th target vector through the binomial crossover as explained earlier.

Selection step

Evaluate the trial vector and update the current generation with the better individual vector.

END FOR

Increase the generation counter $G=G+1$

END WHILE

3 Transportation Dataset

This is Home Interview Survey (HIS) data, which is collected from a survey made in the city of Hyderabad. The main purpose of this project is to understand the present day travel patterns and relate these patterns to the Socio-Economic characteristics of Trip makers [16].

The travel patterns in the form of number of trips performed by each member in a city, from an identifiable location in the city called Origin to another identifiable location called Destination, together with the trip makers socio economic characteristics, is the primary bed block based on which future predictions of travel are made. This information is used in developing travel demand models that will help in predicting future travel patterns for the horizon year. These predicted travel patterns are the main source of information for identifying, planning, locating, designing, justifying various transportation projects. For calibrating the Demand model, base year travel patterns along with their attributes are necessary. For this purpose elaborate travel surveys are organized. The main method of obtaining all these travel attributes from road users is to elicit from them either directly by interviews or obtaining indirectly by phone call or through written reply by mail or e-mail. The principal methods of intercepting the transport users are either at the beginning of trip called Origin end, or at end of trip called Destination end.

In order to present the findings of these surveys it is necessary to translate the information into origin destination matrices stratified by purpose, mode and time of travel etc., for further applications. Since there are innumerable number of Origins and Destinations in the study region, it is not possible to describe each trip from their exact place of origin to exact place of destination. Rather, the study area is divided into small Traffic Analysis Zones (TAZ) or localities to represent a group of houses or group of activities. All those trips that start or end anywhere in the TAZ area, are assumed to be originating at or destined to the centroid of the TAZ. Thus all individual homes and activities are aggregated to reasonable number of representative traffic analysis zones. This finite number of representative origins and destinations enables to generate O-D matrices that are reasonably workable, for computations and model development. However care is to be exercised to ensure that zone sizes are not too big as to distort the travel patterns, or nor too small that the secondary data becomes difficult to obtain and predict for future, or becomes not compatible with transport network. In the present study the data was obtained from Hyderabad city Home Interview Surveys, divided into 147 Traffic Analysis Zones or localities.

In the Urban Transportation Planning package modeling system, travel is expressed as a function of socio economic characteristics of the traveler, activity type and intensity at the destination end of trip and level of service provided by the transport network and modes of travel. The expression would be of the type:

$T_{ijmrt}^n = f(S, L, A)$ Where

T_{ijmrt}^n = Travel made from zone 'i' to zone 'j', by mode 'm' through route 'r', in time 't', for the purpose 'n'

S = socio economic characteristic of trip maker

L = level of service provided by the transport system between 'i' and 'j'

A = activity system variable at zone 'j'

The main issue in our work is to capture the relevant attributes from the dataset given to us and once it is done we need to group the records based on their characteristics to find out some interesting inferences. It may be noted here that the dataset available have no distinct groups specified or in other words it is unsupervised. The main challenge is to determine the clusters of homogeneous records. In our simulation we have first attempted to use classical k-means clustering approach to find clustering by taking 2,4,6 and 8 clusters as input in starting point and trying to minimize the square error . Although we could arrive at finding a cluster of 4 giving minimized square error but the computational time for the large dataset was a bottleneck and the possibility of being trapped in local minima also could not be fully averted.

To alleviate above problems, we have used principal component analysis approach in this work [17] to obtain best relevant features. And after that we have implemented adaptive DE to have Dynamic clustering. In the next section, we present the work on pre-processing of dataset based on relevant feature selection with PCA and followed by the dynamic DCADE algorithm to feature the selected dataset.

3.1 Pre-processing of Dataset

The data collected from Home Interview Surveys contains intercorrelations among the elements of the socio economic vector variable which makes it difficult to construct and interpret any model from them. To detect these relationships one can use projections along different directions defined by a weighted linear combination of variables, or components that maximize the variance subject to being uncorrelated. Principal component analysis is a useful procedure to determine the minimum number of independent dimensions needed to account for most of the variance in the original test data. They do reveal fewer independent dimensions that are required to define the test domain. Factor analysis further improves the solutions offered by principal components by rotating components to positions that are most interpretable. With only few variables, it is easier to search interesting spaces manually by rotating the distribution data. In this paper we have employed varimax solution for rotation of axis. For each factor, varimax rotation also yields high loadings for a few variables that will help in understanding basic trait associated with the factor[18].

4 Dynamic Clustering to HIS Data

DCADE algorithm has been simulated with 26 relevant features of the given dataset. The minimum and maximum numbers of clusters are chosen for dynamic clustering taken from 2 to 10. Accordingly the vectors are initialized in the initial population. To compare with other dynamic clustering algorithms such as DCPSO, GUCK and Classical DE, same range is taken for all approaches. The simulations are done for 10 runs and results are reported in tabular form below (Table 1).

From the table 1 it is clearly evident that the optimum numbers of clusters found are 4 which is in accordance with the domain expert. Our proposed adaptive DE gives the closest results. After going through the data vectors belonging to each cluster as found by DCADE we have analyzed them and concluded with following inferences with regard to the nature of the group.

Table 1. Comparison of Algorithms

Dataset	Transport			
	DCADE	DCPSO	GUCK	Classical DE
Avg no. of Clusters found (std)	4.02(0.001)	8.673(0.003)	14.667(0.019)	0.6723(0.099)
Mean intra Cluster Distance(std)	5.2(0.0469)	10.912(0.3570)	11.392(0.228)	0.912(0.0037)
Mean Inter Cluster Distance(std)	4.72(0.005)	9.653(0.0023)	13.972(0.0257)	0.702(0.032)
XB measure (std)	4.92(0.043)	9.012(0.002)	12.37(0.0656)	0.812(0.0076)

- **Cluster I:** Persons having high Income, having car ownership, persons having fewer dependents, or more earners, houses located away from public transport systems etc. In other words they belong to Prosperous families. One can identify the prosperity by associating Car with the house.

- **Cluster II:** The second class people are those whose incomes are slightly lower than the above class, but with more dependents, or less earners, with slightly less educational standards. They possess at least a Two wheeler like Scooter, Motor cycle, Moped etc. This group can be considered as upper middle class and can be considered as those who own Two wheeler.

- **Cluster III:** The third category of people mostly does not have a vehicle but, may have occasionally Two wheeler, but they prefer to travel by public transport. Their family size is slightly bigger, and this group can be considered as middle income group people. May be considered has no vehicle owning group.

- **Cluster IV:** The fourth category have low educational level, work in private sector, or work in some activity on daily wage basis. Mostly they have a bicycle if they work in fixed time schedule activity, or on contract basis. They have slightly lower type of residences. This group of people can be considered as having Bicycle owning group.

Accordingly these four classes can now be identified by their vehicle ownership levels. Those that are Vehicle owning and those that are No Vehicle owning classes. Among the Vehicle owning group they can be further classified into those having a Car ownership, Two wheeler ownership, Bicycle ownership etc. In other words, the behaviour of these classes of people especially in the context of travel behaviour can be identified on that nomenclature. They are distinctly different from each other.

With these findings, it is possible to visualize the future scenarios, should it be possible to predict the type of employment, the number of earning members in a house hold or their changing patterns in occupational structure, on the changing patterns of traffic composition. The travel demand can be realistically estimated if the population is stratified into four clusters and develop relationships separately rather than on single population.

5 Conclusion

In this exploration dynamic clustering using ADE called DCADE, has been used to obtain the optimal number of clusters at runtime. DCADE algorithm is applied on a real time data known as Home Interview Survey (HIS) data which is related to a Transport Project. Later clusters are formed and analyzed which are in accordance with the domain expert.

References

1. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: On clustering validation techniques. *J. Intell. Inf. Syst.* 17(2/3), 107–145 (2001)
2. Theodoridis, S., Koutroubas, K.: *Pattern Recognition*. Academic, New York (1999)
3. Rosenberger, C., Chehdi, K.: Unsupervised clustering method with optimal estimation of the number of clusters: Application to image segmentation. In: *Proc. IEEE ICPR*, Barcelona, Spain, vol. 1, pp. 656–659 (2000)
4. Lee, C.-Y., Antonsson, E.K.: Self-adapting vertices for mask-layout synthesis. In: Laudon, M., Romanowicz, B. (eds.) *Proc. Model. Simul. Microsyst. Conf.*, San Diego, CA, March 27–29, pp. 83–86 (2000)
5. Schwefel, H.-P.: *Evolution and Optimum Seeking*, 1st edn. Wiley, New York (1995)
6. Fogel, L.J., Owens, A.J., Walsh, M.J.: *Artificial Intelligence Through Simulated Evolution*. Wiley, New York (1966)
7. Bandyopadhyay, S., Maulik, U.: Genetic clustering for automatic evolution of clusters and application to image classification. *Pattern Recognit.* 35(6), 1197–1208 (2002)
8. Omran, M., Engelbrecht, A., Salman, A.: Particle swarm optimization method for image clustering. *Int. J. Pattern Recognit. Artif. Intell.* 19(3), 297–322 (2005)
9. Kennedy, J., Eberhart, R.C.: A discrete binary version of the particle swarm algorithm. In: *Proc. IEEE Conf. Syst., Man, Cybern.*, pp. 4104–4108 (1997)
10. Kennedy, J., Eberhart, R.C.: A discrete binary version of the particle swarm algorithm. In: *Proc. IEEE Conf. Syst., Man, Cybern.*, pp. 4104–4108 (1997)
11. Das, S., Abraham, A.: Automatic Clustering using and Improved Differential Evolution Algorithm. *IEEE Transactions on Systems, Man and Cybernetics* 38(1) (January 2008)
12. Halkidi, M., Vazirgiannis, M.: Clustering validity assessment: Finding the optimal partitioning of a data set. In: *Proc. IEEE ICDM*, San Jose, CA, pp. 187–194 (2001)
13. Xie, X., Benni, G.: A Validity Measure for Fuzzy Clustering. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)* 13(8), 841–847 (1991)
14. Brest, J., Greiner, S., Böškovíc, B., Mernik, M., Zumer, V.: Self adapting control parameters in differential evolution: A comparative study on numerical benchmark problems. *IEEE Trans. Evol. Comput.* 10(6), 646–657 (2006)
15. Zaharie, D.: Critical values for the control parameters of differential evolution algorithms. In: *Proc. 8th Int. Mendel Conf. Soft Comput.*, pp. 62–67 (2002)
16. Teo, J.: Exploring dynamic self-adaptive populations in differential evolution. *Soft. Comput.* 10, 673–686 (2006), doi:10.1007/s00500-005-0537-1
17. Development of Hyderabad Multimodal Suburban Commuter Transportation System, Government of Andhra Pradesh (2004)
18. Anand, S., Sai Hanuman, A., et al.: Use of Data Mining Techniques in understanding Home Interview Surveys Employed for Travel Demand Estimation. In: *International Conference DMIN 2008, World Congress in Computer Science, Los Vegas, USA* (2008)

NLMS Algorithm Based CMA Channel Equalization through an Adaptive MMSE Equalizer

Rangisetty Nirmala Devi¹, Tara Saikumar², and K. Kishan Rao³

¹ Dept of EIE, KITS Warangal, India

² Dept of ECE, CMR Technical Campus Hyderabad, India
nimala123@yahoo.com, tara.sai437@gmail.com

³ Vaagdevi Group of institutions, Warangal, India
prof_kkr Rao@rediffmail.com

Abstract. The adaptive algorithm has been widely used in the digital signal processing like channel estimation, channel equalization, echo cancellation, and so on. One of the most important adaptive algorithms is the NLMS algorithm. We present in this paper an multiple objective optimization approach to fast blind channel equalization. By investigating first the performance (mean-square error) of the standard fractionally spaced CMA (constant modulus algorithm) equalizer in the presence of noise, we show that CMA local minima exist near the minimum mean-square error (MMSE) equalizers. Consequently, CMA may converge to a local minimum corresponding to a poorly designed MMSE receiver with considerably large mean-square error. The step size in the NLMS algorithm decides both the convergence speed and the residual error level, the highest speed of convergence and residual error level.

Keywords: CMA, NLMS, Adaptive MMSE.

1 Introduction

Blind equalization has the potential to improve the efficiency of communication systems by eliminating training signals. Difficulties of its application in wireless communications, however, are due largely to the characteristics of the propagation media - multipath delays and fast fading. The challenge is achieving blind equalization using only a limited amount of data.

A widely tested algorithm is the constant modulus algorithm (CMA). In the absence of noise, under the condition of the channel invertibility, the CMA converges globally for symbol-rate IIR equalizers and fractionally spaced FIR equalizers. It is shown in [9] that CMA is less affected by the ill-conditioning of the channel. However, Ding *et al.* [2] showed that CMA may **converge** to **some** local minimum for the symbol rate FIR equalizer. In the presence of noise, the analysis of convergence of CMA is difficult and little conclusive results are available. Another drawback of CMA is that its convergence rate may not be sufficient for fast fading channels. Another approach to the blind equalization is based on the blind channel estimation. Some of the recent eigen structure-based channel estimations require a relatively smaller data size comparing with higher order statistical methods. However

the asymptotic performance of these eigen structure-based schemes is limited by the condition of the channel [12, 13]. Specifically, the asymptotic normalized mean square error (ANMSE) is lower bounded by the condition number of the channel matrix. Unfortunately, frequency selective fading channels with long multipath delays often result in ill conditioned channel matrices. The key idea of this paper is to combine the approach based on minimizing the constant modulus cost and that based on matching the second-order cyclostationary statistics. The main feature of the proposed approach is the improved convergence property over the standard CMA equalization and the improved robustness for ill-conditioned channels.

2 Blind Channel Equalization and Types

The field of blind channel equalization has been existence for a little over twenty years. Research during this time has centered on developing new algorithms and formulating a theoretical justification for these algorithms. Blind channel equalization is also known as a self-recovering equalization. The objective of blind equalization is to recover the unknown input sequence to the unknown channel based solely on the probabilistic and statistical properties of the input sequence. The receiver can synchronize to the received signal and to adjust the equalizer without the training sequence. The term blind is used in this equalizer because it performs the equalization on the data without a reference signal. Instead, the blind equalizer relies on knowledge of the signal structure and its statistic to perform the equalization.

1. Blind signal is the unknown signal which would be identified in output signal with accommodated noise signal at receiver.
2. Channel equalization uses the idea & knowledge of training sequences for channel estimation where as Blind channel equalization doesn't utilizes the characteristics of training sequences for frequency and impulse response analysis of channel.
3. Blind Channel Equalization differs from channel equalization and without knowing the channel characteristics like transfer function & SNR it efficiently estimate the channel and reduces the ISI by blind signal separation at receiver side by suppressing noise in the received signal.

3 CMA – Constant Modulus Algorithm

In digital communication, equalizer was designed to compensate the channel distortions, through a process known as equalization. There are two types of equalization which are: 1) Trained equalization, 2) Blind (self recovering) Equalization.

Blind equalization finds important application in data communication system. In data communications, digital signals are generated and transmitted by the sender through an analog channel to the receiver. Linear channel distortion as a result off line limited channel bandwidth, multipath and fading is often the most serious distortion in digital communication system. Blind equalization improves system bandwidth efficient by avoiding the use of training sequence. The linear channel distortion, known as the Inter symbol interference (ISI), can severely corrupt the transmitted signal and make it difficult for the receiver to directly recover the transmitted data.

Channel equalization and identification has proven to be an effective means to compensate the linear distortion by removing much of the ISI.

Channel Equalization

A typical communication system design involves first passing the signal to be transmitted through a whitening filter to reduce redundancy or correlation and then transmitting the resultant whitened signal. At the receiver, the recorded signal is passed through the inverse whitening filter and the original signal is thus restored. However, the channel will affect the transmitted signal because of a) Channel noise b) Channel dispersion leading to inter symbol interference. For example, by reflection of the transmitted signal from various objects such as buildings in the transmission path, leading to echoes of the transmitted signal appearing in the receiver. Therefore, it is necessary to pass the received signal through a so called equalizing filter to undo the dispersion effect as shown in figure 2 below. Equalization compensates for Inter symbol Interference (ISI) created by multi path within time dispersive Channel message signal whitening signal receiver.

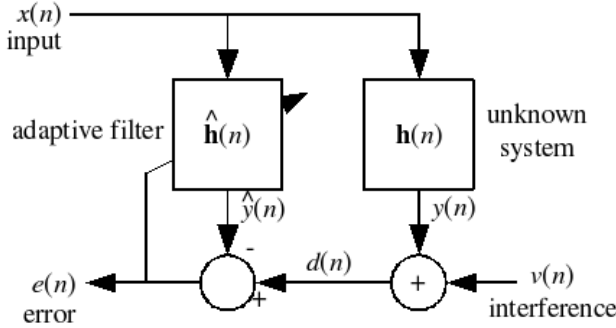
Blind Channel Equalization

The field of blind channel equalization has been existence for a little over twenty years. Research during this time has centered on developing new algorithms and formulating a theoretical justification for these algorithms. Blind channel equalization is also known as a self-recovering equalization. The objective of blind equalization is to recover the unknown input sequence to the unknown channel based solely on the probabilistic and statistical properties of the input sequence. The receiver can synchronize to the received signal and to adjust the equalizer without the training sequence. The term blind is used in this equalizer because it performs the equalization on the data without a reference signal. Instead, the blind equalizer relies on knowledge of the signal structure and its statistic to perform the equalization. A natural question from for direct adaptive equalization with training is, "How can we adapt our filter F , without the use of a training signal?". Figure 2 shows such a system. There has been extensive research on this subject for single user applications as well as multi-user applications. The Constant Modulus Algorithm is one such algorithm employed for the blind adaptation problem.

4 NLMS Algorithm

Usually, the adaptive algorithm consists of a transfer filter for processing the input single and an algorithm unit for update the transfer filter's coefficients. $x(n)$ is the input signal; $w(n) = [w_0, w_1, w_2, \dots, w_l]$ is the vector of the transfer filter's coefficients; $d(n)$ is the desired output of the transfer filter; $y(n)$ is the output of the transfer filter; $e(n)$ is the error value, and it can be written as:

$$e(n) = d(n) - \hat{y}(n) \quad (1)$$



The Adaptive algorithm unit represents some algorithm to update the coefficients of the transfer filter. For LMS algorithm, the method to update the coefficients of the transfer filter is given as follows:

$$w(n) = w(n + 1) + \mu^* x(n) * e(n) \tag{2}$$

μ , is the step of LMS algorithm.

The main drawback of the "pure" LMS algorithm is that it is sensitive to the scaling of its input $x(n)$. This makes it very hard (if not impossible) to choose a learning rate μ that guarantees stability of the algorithm. The *Normalised least mean squares filter* (NLMS) is a variant of the LMS algorithm that solves this problem by normalising with the power of the input. The NLMS algorithm can be summarised as:

Parameters: p = filter order μ = step size Initialization: $\hat{h}(0) = 0$
 Computation: For $n = 0, 1, 2, \dots$

$$X(n) = [x(n), x(n-1), \dots, x(n-p+1)]^T$$

$$e(n) = d(n) - \hat{h}^H(n) X(n) \quad \hat{h}(n+1) = \hat{h}(n) + \frac{\mu e^*(n) X(n)}{X^H(n) X(n)}$$

5 Adaptive MMSE Equalizer

The Sampled signal after MMSE Equalizer can be expressed in matrix form as

$$s(i) = w^H y(i) \tag{3}$$

Where $y(i) = H^T(i)s(i) + n(i)$, (4)

M is the length of the MMSE Equalizer: $w = [w_1, w_2, w_3, w_4, w_5, \dots, w_M]^T$ is the equalizer coefficients vector; Then the error signal $e(i)$ is given by

$$e(i) = d(i) - \hat{s}(i) \tag{5}$$

where $d(i)$ is the desired response. For MMSE equalizer, $d(i) = s(i + D)$, D is a time delay parameter which is $L + 1$ usually. The MMSE criterion is used to derive the optimal equalizer coefficients vector w :

$$w = \underset{w}{\text{minimize}} E\{|e|^2\} \tag{6}$$

We make the assumption that signal $s(i)$ and noise $n(i)$ are independent identity distribution stochastic.

Variable and uncorrelated each other, then the equalizer coefficients vector w can be expressed as[2]:

$$w = (H^H H + \frac{1}{SNR} I)^{-1} H^H \delta_D \tag{7}$$

Where $\delta_D = [0 \dots 1_D; 0 \dots 0]_{1 \times (L+M-1)}^T$, $SNR = \frac{\sigma_s^2}{\sigma_n^2}$ denotes the signal noise

ratio I is $M \times M$ identity matrix.

To reduce the complexity caused by matrix inversion of ideal MMSE equalizer, we propose an adaptive MMSE equalizer algorithm. In code-multiplexed pilot CDMA systems, conventional adaptive equalizer is difficult to implement for lack of reference signal. In this paper, the steepest descent method[4] is used to derive adaptive equalizer algorithm in code-multiplexed pilot CDMA systems.

According to Eqn.3 and Eqn.5, the mean square error (MSE) J can be expressed as

$$J(w) = E[e(i)e(i)^*] = \sigma_s^2 - w^H p - p^H w + w^H R w \tag{8}$$

where autocorrelation matrix $R = E[y(i)y^H(i)]$; cross-correlation vector $p = E[y(i)d^*(i)]$, σ_s^2 denotes the signal power; $(.)^*$ represents conjugate operation. Because the wireless channel is time-varying, the equalizer coefficients vector w must be updated real time. Conventional adaptive algorithm requires reference signal $d(i)$, while in the downlink of code-multiplexed pilot CDMA systems, $d(i)$ is difficult to distill. To resolve this problem, the steepest decent method is used. From Eqn.8, the gradient vector is

$$\frac{\partial J(w)}{\partial w} = -2p + 2Rw \tag{9}$$

then the equalizer coefficients updating equation is

$$w(i+1) = w(i) + 2\mu[p - Rw(i)] \tag{10}$$

where parameter μ is a positive real-valued constant which controls the size of the incremental correction applied to the equalizer coefficients vector.

For the autocorrelation matrix:

$$\begin{aligned}
 R &= E[y(i)y^H(i)] \\
 R &= E[s(i)s^H(i)]\{H^H(i)H(i)\}^T + E[n(i)n^H(i)] \\
 R &= \sigma_s^2 \{H^H(i)H(i)\}^T + \sigma_n^2 I
 \end{aligned} \tag{11}$$

the cross-correlation vector

$$\begin{aligned}
 p &= E[y(i)d^*(i)] = E[(H^T(i)s(i) + n(i))s^*(i - D)] \\
 p &= \sigma_s^2 H^T(i)\delta_D
 \end{aligned} \tag{12}$$

From Eqn.7,8,9, we can obtain the time recursive equation of MMSE equalizer by:

$$w(i+1) = w(i) + 2\mu\sigma_s^2 [H^T(i)\delta_D - \{H^H(i)H(i)\}^T w(i)] + \frac{1}{SNR} Iw(i) \tag{13}$$

As can be seen from Eqn.13, the updating process avoids the matrix inversion operation. On the other hand, the updating process abstains the requirement to store the autocorrelation matrix $\mathbf{R}(i)$ and only the equalizer coefficients vector of last time is needed. From Eqn.13 we know, the channel convolution matrix $\mathbf{H}(i)$ is required to update the equalizer coefficients vector.

For CMA, channel response can be estimated through code-multiplexed pilot. In this paper, the low complexity sliding-window method is used to estimate the channel coefficients, which can be expressed as

$$\hat{\beta}_l(i) = \frac{1}{2\sqrt{\alpha pw(i+1)T_s}} \int_{\tau_l + (i - \frac{w}{2})T_s}^{\tau_l + (i + \frac{w}{2})T_s} y(t)c_p^*(t - \tau_l) dt \tag{14}$$

where $\hat{\beta}_l(i)$ is estimation of the complex gain of l -th path; w is the length of sliding-window in symbols and should be selected properly according to the varying speed of the channel.

6 Experimental Results

The performances for the Adaptive MMSE and adaptive CMA algorithm through and NLMS algorithm is experimental performed with accurate figures from 1-8.

The performances of the channel estimation is an analyze in such a way that transmitted bits and receiver bits, Equalizers and very important task that is Convergences.

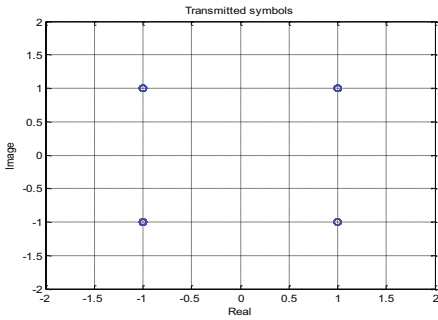


Fig. 1. Transmitter side of Adaptive CMA

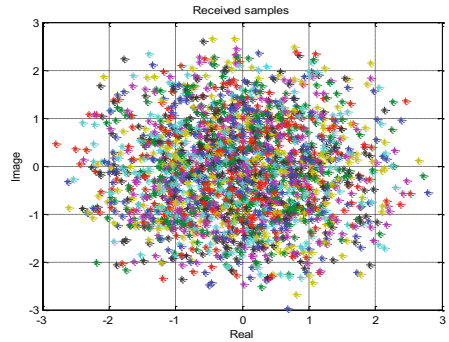


Fig. 2. Receiver side of Adaptive CMA

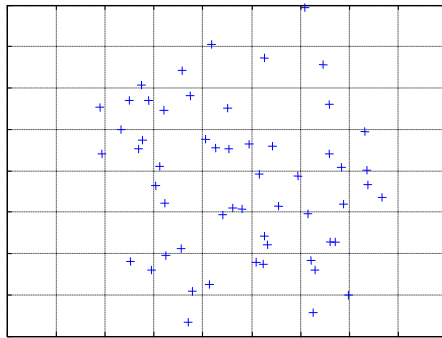


Fig. 3. Adaptive CMA Equalizer

The first figures from 1-4 are obtained for Adaptive CMA Equalizer, with more efficient for equalization and convergences. Secondly from figure 5-8 are obtained for an Adaptive MMSE equalizer through NLMS algorithm. The efficient of equalization and convergences is too good. The time complexity is very less and more efficient for advance communication systems.

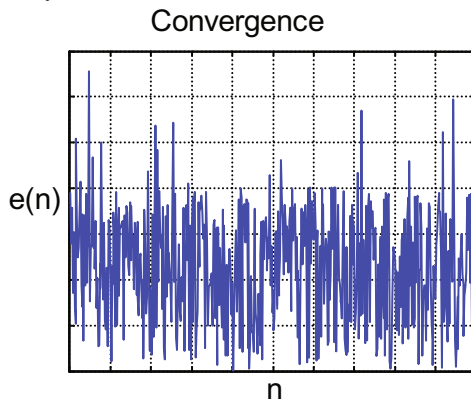


Fig. 4. Convergence of Adaptive CMA equalizer

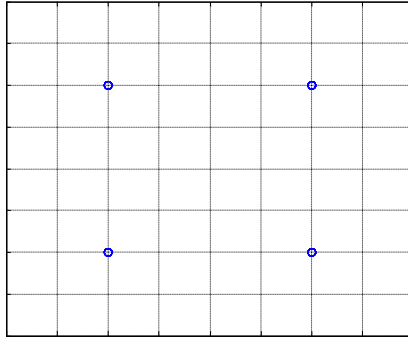


Fig. 5. Transmitter side of Adaptive MMSE Equalizer

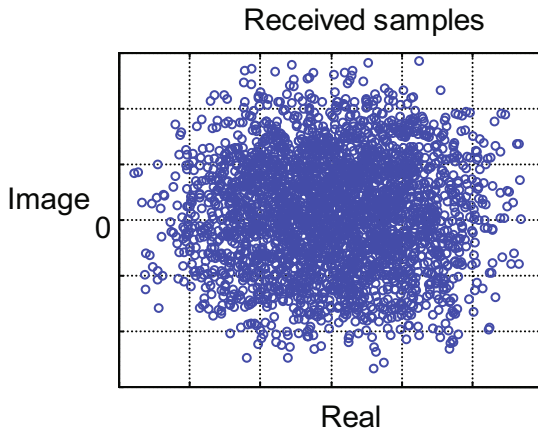


Fig. 6. Receiver side of Adaptive MMSE Equalizer

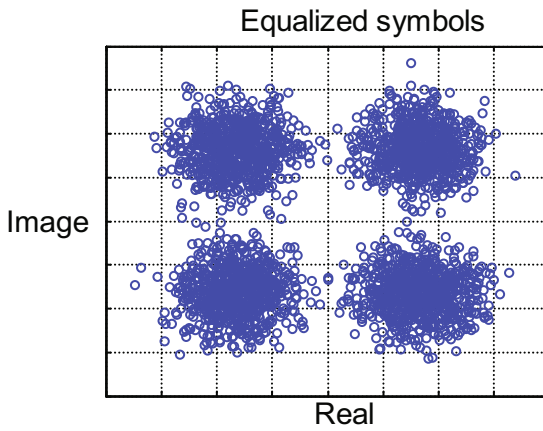


Fig. 7. Adaptive MMSE Equalizer through NLMS

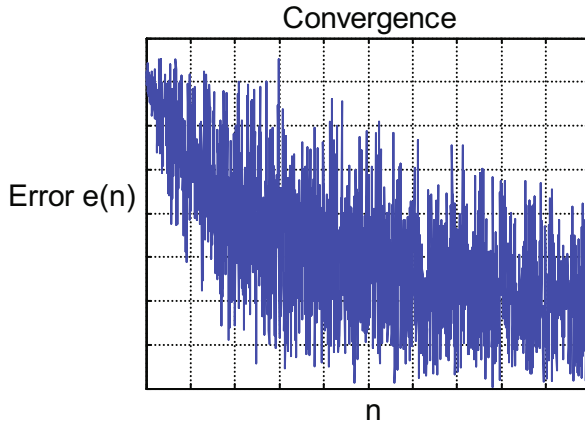


Fig. 8. Convergence of Adaptive MMSE Via NLMS algorithm

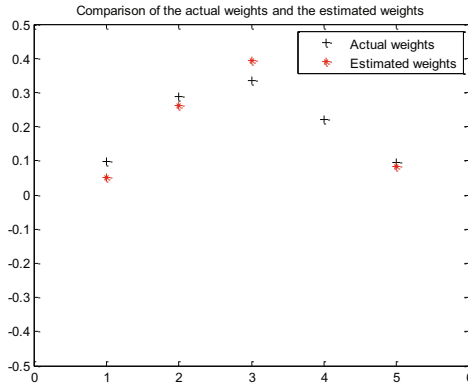


Fig. 9. Comparison of the actual weight and estimated weight

7 Conclusions

In this paper, the implementation of NLMS Algorithm with Adaptive MMSE gives more accurate results compare to with LMS Algorithm with Adaptive MMSE that satisfy the performance requirement in high data rate transmission, while ideal MMSE equalizer is difficult to real-time implement because its large computational complexity, a low complexity adaptive MMSE equalizer algorithm is proposed. In future conclusion, the proposed low complexity adaptive MMSE equalizer in code-multiplexed CDMA system can be proposed and this system has better practical application value.

References

- [1] Krauss, T.P., Zoltowski, M.D., Leus, G.: Simple MMSE equalizers for CDMA downlink to restore chip sequence: Comparison to Zero-Forcing and Rake. In: ICASSP, vol. 5, pp. 2865–2868 (2000)
- [2] Hooli, K., Latva-aho, M., Juntti, M.: Multiple access interference suppression with linear chip equalizers in WCDMA downlinkreceivers. In: General Conference (Part A), GLOBECOM, pp. 467–471 (December 1999)
- [3] Mailaender, L.: Low-complexity implementation of CDMA downlink equalization. *3G Mobilec Communication Technologies* 477, 396–400
- [4] Haykin, S.: *Adaptive Filter Theory*, 3rd edn. Prentice Hall (1996)
- [5] Golub, G.H., Van Loan, C.F.: *Matrix Computation*, 3rd edn. The Johns Hopkins University Press (1996)
- [6] Shynk, J.: Frequency-domain and multirate adaptive filtering. *IEEE Signal Processing Magazine* 9, 14–39 (1992)
- [7] Godard, D.N.: Self-recovering equalization and carrier tracking in two-dimensional data communication systems. *IEEE Trans. on Communications*
- [8] Fijalkow, I., Manlove, C.E., Johnson Jr., C.R.: Adaptive fractionally spaced blind CMA equalization: Excess MSE. *IEEE Trans. on Signal Processing* 46(1), 227–231 (1998)
- [9] Ding, Z., Kennedy, R.A., Anderson, B.D.O., Johnson Jr., C.R.: Ill-convergence of godard blind equalizers in data communication systems. *IEEE Trans. On Communications* 39
- [10] Johnson Jr., C.R., Dasgupta, S., Sethares, W.A.: Averaging analysis of local stability of a real constant modulus algorithm adaptive filter. *IEEE*
- [11] Brown, D.R., Schniter, P.B., Johnson Jr., C.R.: Computationally efficient blind equalization. In: 35th Annual Allerton Conference on Communication, Control, and Computing (September 1997)
- [12] Casas, R.A., Johnson Jr., C.R., Kennedy, R.A., Ding, Z., Malamut, R.: Blind adaptive decision feedback equalization: A class of channels resulting in illconvergence from a zero initialization. *International Journal on Adaptive Control and Signal Processing Special Issue on Adaptive Channel Equalization*
- [13] Johnson Jr., C.R., Anderson, B.D.O.: Godard blind equalizer error surface characteristics: White, zeromean, binary source case. *International Journal of Adaptive Control and Signal Processing* 9, 301–324
- [14] Nirmala Devi, R., Saikumar, T., Kishan Rao, K.: Adaptive MMSE Equalizer through LMS algorithm based CMA Channel Equalization. In: The Thrid International Conference on Network and Communications (NetCom-3.0), (accepted , Publish in LNICST, January 2-4, 2012)

A Comprehensive Study of Particle Swarm Based Multi-objective Optimization

Samantula Mohankrishna¹, Divya Maheshwari², P. Satyanarayana³,
and Suresh Chandra Satapathy⁴

¹ Dept of IT, GIT, Gitam University, India
smkrishna@ieee.org

² IME Sahibabad-Ghaziabad
maheshwari.divya84@gmail.com

³ Vizag Steel, Vishakapatnam

s_paramata@vizagsteel.com

⁴ Dept of CSE, Anil Neerukonda Institute of Technology and Science, India
sureshsatapathy@ieee.org

Abstract. Recently there has been a growing interest in evolutionary multiobjective optimization algorithms which combines two major disciplines: evolutionary computation and the theoretical frameworks of multicriteria decision making. This paper presents a comprehensive study of Multi-Objective Optimization (MOO) with Particle Swarm Optimization (PSO). Different suggestions of various researchers have been compiled to give a first-hand information of PSO based MOO. It is found that no single approach is superior. Rather, the selection of a specific method depends on the type of information that is provided in the problem, the user's preferences, the solution requirements and the availability of software.

Keywords: Multi objective, Particle swarm optimization, PSO, Social networks, Swarm theory, Swarm dynamics.

1 Introduction to Multi-objective Optimization (MOO) and PSO

Optimization plays a major role in any branch of engineering. It is the mechanism by which the maximum or minimum value of a function or process is computed. This is used to optimize efficiency, production, profit or some other measure etc. Optimization can refer to either minimization or maximization; maximization of a function f is equivalent to minimization of the opposite of that function. In multi-objective optimization a collective of objective functions are systematically and simultaneously optimized. Even though many works have been done in multi-objective optimization areas still it is a very important research topic both for scientists and engineers, because there are still many open questions in this area. The interesting fact of this is that there is no accepted definition of "optimum" as in single-objective optimization. Hence it is difficult to even compare the results of one method to another's because, normally, the decision about the "best" answer corresponds to the so-called (human) decision maker.

Multi-objective optimization originally grew out of three areas: economic equilibrium and welfare theories, game theory, and pure mathematics. More precisely, multiobjective problems (MOPs) are those problems where the goal is to optimize k objective functions simultaneously. This may involve the maximization of all k functions, the minimization of all k functions or a combination of maximization and minimization of these k functions. Defining multiple objectives often gives a better idea of the task. In contrast to the plethora of techniques available for single-objective optimization, relatively few techniques have been developed for multiobjective optimization. In single objective optimization, the search space is often well defined. As soon as there are several possibly contradicting objectives to be optimized simultaneously, there is no longer a single optimal solution but rather a whole set of possible solutions of equivalent quality. When we try to optimize several objectives at the same time the search space also becomes partially ordered. To obtain the optimal solution, there will be a set of optimal trade-offs between the conflicting objectives. A multiobjective optimization problem is defined by a function f which maps a set of constraint variables to a set of objective values. An optimal solution is the solution that is not dominated by any other solution in the search space. Such an optimal solution is called Pareto optimal and the entire set of such optimal trade-offs solutions is called Pareto optimal set. As evident, in a real world situation a decision making (trade-off) process is required to obtain the optimal solution. Even though there are several ways to approach a multiobjective optimization problem, most work is concentrated on the approximation of the Pareto set.

Particle Swarm Optimization belongs to the field of Swarm Intelligence and Collective Intelligence and is a sub-field of Computational Intelligence. The idea behind this approach is to simulate the movements of a group (or population) of birds which aim to find food. The approach can be seen as a distributed behavioral algorithm that performs (in its more general version) multidimensional search. This approach is more suited to continuous variable problems. One major characteristic of PSO is it provides many configuration parameters, which allow the algorithm to be adjusted to various problem landscapes. In PSO individuals are allowed benefit from their past experiences whereas in an evolutionary algorithm, normally the current population is the only “memory” used by the individuals. PSO has been successfully used for both continuous nonlinear and discrete binary optimization [1, 3, 5, 6].

2 Particle Swarm Multiobjective Versions

In the last few decades, a variety of PSO techniques for handling multiple objectives have been published in various literatures. Few of the efficient techniques will be reviewed next.

Moore and Chapman [6] presented a Pareto dominance algorithm in an unpublished document. The authors highlighted the value of performing both an cognitive component and a social component (group search). However, the authors did not endorse any scheme to balance diversity. Liew and Ray [8] proposed the swarm metaphor algorithm, which uses Pareto dominance and combines concepts of evolutionary techniques with the particle swarm. This approach used crowding to maintain diversity and a multilevel filter to handle constraints. The authors adopt the constraint and

objective matrices proposed in some of their previous research [9] to implement the approach. Contrasting with the previous authors proposals, Parsopoulos and Vrahatis [10] adopted an aggregating function and are implemented through three types of approaches: a conventional linear aggregating function, a dynamic aggregating function and the bang bang weighted aggregation approach [11]) for their multi-objective PSO approach. In modern work, Parsopoulos [12] studied a parallel version of the Vector Evaluated Particle Swarm (VEPSO) method for multiobjective problems. VEPSO is a multi-swarm variant of PSO, which is inspired on the Vector Evaluated Genetic Algorithm (VEGA) [13]. In VEPSO, evaluation of each swarm is done using only one of the objective functions of the problem under consideration, and the information that is possessed for the objective function is communicated to the remaining swarms through the sharing of their best experience. In 2002 Hu and Eberhart [14] proposed an approach called “dynamic neighborhood”, where only one objective is optimized at a time using a scheme similar to lexicographic ordering. Discussed mainly about Dynamic neighbors, new pBest updating strategy and one dimension optimization. In further work, Hu [15] adopt a secondary population (called “extended memory”) and proposed some further improvements to their dynamic neighborhood PSO approach. Fieldsend and Singh [16] has come out with an approach that uses unconstrained elite archive, where special data structure “dominated tree technique “ is adopted to store the nondominated population found along the search process. The primary population interacts with the archives in order to define local guides. Their approach also uses a “turbulence” operator that is basically a mutation operator that acts on the velocity value used by PSO. Coello Coello and Salazar Lechuga [30] and Coello Coello et al. [19] proposed an approach based on the idea of having a global repository in which every particle deposits its flight experiences after each flight cycle. Additionally, the updates to the repository are performed considering a geographically- based system defined in terms of the objective function values of each individual; this repository is used by the particles to identify a leader that will guide the search. The approach also uses a mutation operator that acts both on the particles of the swarm, and on the range of each design variable of the problem to be solved.

In more recent work, Toscano Pulido and Coello Coello [19] use the concept of Pareto dominance to determine the flight direction of a particle. In order to have a better distribution of solutions in decision variable space the authors adopted various clustering techniques to divide the population of particles into several swarms. PSO algorithm is executed with each sub-swarm and at later point, the different sub-swarms exchange information (the leaders of each swarm are migrated to different swarm in order to vary the selection pressure. Also, this approach does not use an external population since elitism in the current case is an emergent process derived from the migration of leaders. For solving water quality problems the authors Baltar and Fontane [20] adopted a variation of Coello’s approach [19].The fundamental change is that Baltar and Fontane [20] did not use the adaptive grid of the original proposal, but instead, they calculate in objective function space. The repository in this case is a simple archive that stores the nondominated solutions found along the evolutionary process, but it does not work as a diversity-preserving mechanism, as in the original proposal [19]. An interesting aspect of this work is that the algorithm is implemented in a spreadsheet format using Microsoft Excel and Visual Basic. Tayal [21]

implemented PSO with a linear aggregating function to solve various engineering optimization problems, including the design of: (1) a 2 degrees-of freedom spring mass system, (2) a coil compression spring, (3) a two-bar truss, (4) a gear train and (5) a welded beam. Using external penalty and the weights for the linear aggregating function Constraints are handled. Summarizing, this work illustrates the most straightforward way of using PSO as a single-objective optimizer to solve multiobjective optimization problems.

Mostaghim and Teich [22] propose a sigma method in which the local best guides each particle that are adopted to improve the convergence and diversity of a PSO approach used for multiobjective optimization. They also use a “turbulence” operator, but implemented on a decision variable space. The idea of the sigma method is similar to compromise programming. The use of the sigma values increases the selection pressure of PSO (which was already high). This may cause premature convergence in some cases. In more recent work, Mostaghim and Teich [22] propose a novel method called covering MOPSO (cvMOPSO). The proposed method works in two phases. In phase 1: a MOPSO algorithm is implemented with a restricted archive size and the goal is to obtain a good approximation of the Pareto-front. In phase 2, the nondominated solutions obtained from phase 1 are considered as the input archive of the cvMOPSO. The particles in the population of the cvMOPSO are divided into sub swarms around each nondominated solution after the first generation. Li [24] proposes an approach that incorporates the main mechanisms of the NSGA-II [26] to the PSO algorithm. This approach combines the population of particles and all the personal best positions of each particle, and selects the best particles among them to conform the next population. It also selects the leaders randomly from the leaders set among the best of them, based on two different mechanisms: a niche count and a crowding distance. In more recent work, Li [26] proposes the maximinPSO, which uses a fitness function derived from the maximin strategy to determine Pareto domination. The author shows that one advantage of this approach is that no additional clustering or niching technique is needed, since the maximin fitness of a solution can tell us not only if a solution is dominated or not, but also if it is clustered with other solutions, i.e., the approach also provides diversity information.

Srinivasan and Hou [28, 29] propose an approach, called Particle Swarm Inspired Evolutionary Algorithm which is a hybrid between PSO and an evolutionary algorithm. The authors argue that the traditional PSO equations are too restrictive when applied to multiconstrained search spaces. Thus, they propose to replace the PSO equations with the so-called self-updating mechanism, which emulates the workings of the equations. The approach uses a memory to store the elite particles and does not use a recombination operator. Zhang et al. [31] propose an approach that attempts to improve the selection of gbest and pbest when the velocity of each particle is updated. For each objective function, there exists both a gbest and a pbest for each particle. In order to update the velocity of a particle, the algorithm defines the gbest of a particle as the average of the complete set of gbest particles. Analogously, the pbest is computed using either a random choice or the average from the complete set of pbest values. This choice depends on the dispersion degree between the gbest and pbest values of each particle. Zhao & Cao [36] propose a multi-objective particle swarm optimizer based on Pareto dominance. This is very similar to the proposal of Coello and Lechuga [30], since it adopts a gbest topology. However, this approach maintains not one

but two repositories additionally to the main population: one keeps the global best individuals found so far and the other one keeps a single local best for each member of the swarm. A truncated archive is adopted to store the nondominated solutions found along the evolutionary process. Baumgartner et al. [32] propose an approach which uses weighted sums (i.e., linear aggregating functions) to solve multiobjective optimization problem. In this approach, the swarm is equally partitioned into n sub-swarms, each of which uses a different set of weights and evolves into the direction of its own swarm leader. The approach adopts a gradient technique to identify the Pareto optimal solutions.

Chow and Tsui[33] propose an autonomous agent response learning algorithm. The authors propose to decompose the award function into a set of local award functions and, in this way, to model the response extraction process as a multiobjective optimization problem. A modified PSO called “Multi-Species PSO” is introduced by considering each objective function as a species swarm. A communication channel is established between the neighboring swarms for transmitting the information of the best particles, in order to provide guidance for improving their objective values. Mahfouf et al. [35] present an enhancement of the original PSO algorithm which is aimed to improve the performance of this heuristic in multi-objective optimization problems. The approach is called the Adaptive Weighted PSO (AWPSO) algorithm, and its main idea is to modify the velocity by including an acceleration term which increases with the number of iterations. This aims to enhance the global search ability of the algorithm towards the end of the run thus helping the approach to escape from local optima. A weighted aggregating function is also used to guide the selection of the personal and global best leaders. The authors use dynamic weights to generate different elements of the Pareto optimal set. A nondominated sorting scheme is adopted to select the particles from one iteration to the next one. The approach was applied to the design of heat treated alloy steels based on data-driven neural-fuzzy predictive models.

Krami [41] use the MOPSO proposed in [19] to solve reactive power planning problems in which two objectives are minimized: (1) cost and (2) active power losses. This problem has constraints related to the acceptable voltage profiles at each node, which are treated using a death penalty approach (i.e., solutions not satisfying the voltage constraints are discarded). S. Dehuria, S.B. Chob[37], proposed a multi-objective Pareto based particle swarm optimization (MOPPSO) to minimize the architectural complexity and maximize the classification accuracy of a polynomial neural network (PNN). Classification using PNN is explained as a multi-objective problem rather than as a single objective one. Measures like classification accuracy and architectural complexity used for evaluating PNN based classification can be thought of as two different conflicting criterions. Authors has enlighten the use of MOPPSO technique using the two metrics as the criteria of classification problem in finding out a set of non-dominated solution with less complex PNN architecture and high classification accuracy. Shafiq Alam, Gillian Dobbie and Patricia Riddle[38] propose a new generation evolution and swarm intelligence based clustering algorithm called evolutionary particle swarm optimization EPSO-clustering algorithm which represents each cluster with a single particle instead of representing whole clustering solution by individual particles. EPSO-clustering is a combination of two techniques; self organization of the swarm and evolution of the particles at attribute level during a specified generation. Initially a swarm is taken as a clustering solution

and then different particles are merged to optimize the swarm size into an optimal number of particles each representing an individual cluster centroid with associated data vectors.

Juan, Jos, Nebro1 and Carlos A. Coello Coello[39] discussed the search capabilities of six representative state-of-the-art MOPSOs, namely, NSPSO, SigmaMOPSO, OMOPSO, AMOPSO, MOPSO_{pd}, and CLMOPSO. Additionally they propose a new MOPSO algorithm, called SMPSO, characterized by including a velocity constraint mechanism, it is found that SMPSO shows a promising behavior on those problems where the other algorithms fail. Nor ,Mohamad, Ammar[43] presents a study of PSO for constrained optimization problems, discussed the categorization of evolutionary algorithm optimization methods for constrained problems into four types :1) Preserve feasibility of Solutions 2)Penalty function 3) Differentiate the feasible and infeasible solutions 4) Hybrid methods. Discussed various multiobjective problems for PSO. Li Ransikarn Esraa [42] presents a multi-objective diversity guided Particle Swarm Optimization approach named MOPSO-AR which increases diversity performance of multi-objective Particle Swarm optimization by using Attraction and Repulsion (AR) mechanism. AR mechanism uses a diversity measure to control the swarm. Using the approach helps to overcome the problem of premature convergence. AR mechanism integrated with crowding distance computation and mutation operator maintains the diversity of non-dominated set in external archive. Authors demonstrate that the proposed approach is highly competitive in distribution of non-dominated solutions but still keeps convergence towards the Pareto front.

L. Benameur, J.Alami, A. El Imrani[44]proposes a hybrid multiobjective particle swarm approach called Fuzzy Clustering Multiobjective Particle Swarm Optimizer (FC-MOPSO). The model discusses the uses of a fuzzy clustering technique which helps in improving of solution for better distribution in decision variable space by dividing the whole swarm into subswarms In FC-MOPSO, the concept of migration concept is performed in order to interchange information between different subswarms and ensure their diversity. Bin XU Jing YU* YouGan ZHU[45] discusses the idea of escalating strategy which helps to re-generate the whole evolutionary population which results in a new population that is significantly better from the previous values Through this approach the performance on global convergence can be enhanced, and premature can be avoided idea. Yunxia Pei[46]discusses the usability of scheduling technique against two cases: Failure without self-adaptation and No Failure. By the proposed algorithm, the authors claims that not only opportunistic placement of workflow tasks is possible but also significant performance gains are achievable. Moayed Daneshyari, Gary G. Yen[47] introduces a cultural framework which adapts the individual flight parameters of the mutated particles in a MOPSO, namely momentum and personal and global accelerations, The implementation of the proposed algorithm on benchmark test functions shows that the movement of the individual particle using the adapted parameters assists the MOPSO. Junwan Liu Junwan Liu, Yiming Chen [48] discusses the how the development of DNA microarray technology make it very possible to study the transcriptional response of a complete genome to different experimental conditions. A Biclustering technique is shown that is used to analysis those gene expression data. A novel dynamic multi-objective particle swarm optimization biclustering(DMOPSOB) algorithm is discussed here which

is to mine coherent patterns from microarray data. Li Zhongkai, Zhu Zhencai, Zhang Huiqin[49] proposes , a crowding distance sorting based multiobjective particle swarm optimization algorithm (DSMOPSO. With the elitism strategy, the evolution of the external population is achieved based on individuals’ crowding distance sorting by descending order, to delete the redundant individuals in the crowding area. The results show that it outperformed NSGA-II and SPEA2 in the convergence and diversity characteristics of Pareto optimal front. Hsing Hung Lin[50] discusses how a particle swarm optimization (PSO) technique can address open-shop scheduling problems with multiple objectives. This is achieved through modifying the particle position representation, particle velocity, and particle movement to consider the essentially discrete nature of scheduling problems. The proposed algorithm was tested using two benchmark problems to evaluate its performance. The authors conclude that the algorithm performed better when only one swarm was used for all three objectives compared to the case where the swarm was divided into three sub-swarms for each objective.

Gary G. Yen and Wen Fung Leong[51] discussed about functionalities of the the multiobjective constraint handling framework and also about how to update the personal best archive which are designed to encourage finding feasible regions and convergence towards the Pareto front Jintao Yao, Bo Yang, Mingwu Zhang, Yuyan Kong[52].proposes a method of information-sharing by offering particle the predation escaping behavior in order to provide the necessary selection pressure to propel the population moving towards the true Pareto front. The results show that the modified NSPSO can find out the better Pareto Front. Murilo R. Pontes, Fernando B. Lima Neto, Carmelo J. A. Bastos-Filho[53]. discussed the incorporation of auto-adaptation capability in a cooperative Particle Swarm Optimization algorithm, called Clan Particle Swarm Optimization. Carmelo J. A. Bastos-Filho, P’ericles B. C. Mirand[54] discussed about. MOPSOCDR approaches that selects the social and the cognitive leaders based on the crowding distance. Here they proposed a MOPSO with two distinct operation modes. The algorithm changes the operation mode based on the evaluation of the External Archive.

Table 1. Multi-Obejective PSO Variants

	Author (References)	Year	Approach	Application
1	Moore and Chapman [6]	1999	Pareto dominance	-
2	Ray and Liew [8]	2002	swarm metaphor Pareto dominance crowding	Handling Constraints
3	Parsopoulos and Vrahatis [10]	2002	Vector Evaluated	-
4	Fieldsend and Singh [16]	2002	Unconstrained, Elite archive, Dominated tree, turbulence	
5	Hu and Eberhart [13,14]	2003	dynamic neighborhood lexicographic ordering	-

Table 1. (continued)

6	Tayal [21]	2003	aggregating function, external penalty	
7	Mostaghim and Teich [22]	2003	sigma method, "turbulence" operator	Clustering
8	Mostaghim and Teich [23]	2003	"{turbulence" operator	decision variable space
9	Li [24]	2003	a niche count and a crowding distance. Pareto domination	-
10	Zhang et al. [31]	2003	Gbest, pbest	-
11	Coello Coello and Salazar Lechuga [19]	2004	mutation operator	-
12	Toscano Pulido and Coello Coello [18]	2004	Pareto dominance	Clustering
13	Baumgartner et al. [32]	2004	weighted sums, gradient technique	-
14	Chow and Tsui [33]	2004	award function	-
15	Mahfouf [35]	2004	weighted aggregating function	Design of heat treated alloy steels
16	Srinivasan and Hou [28,29]	2005	self-updating mechanism	-
17	Ho . [34]	2005	a "craziness" operator, roulette selection mechanism	
18	Baltar and Fontane [20]	2006	a roulette wheel selection	water quality problems
19	Krami et al. [41]	2006	death penalty approach	Reactive power Planning
20	S. Dehuria, S.B. Chob[37]	2008	Pareto based swarm optimization	Classification
21	Shafiq Alam, Gillian Dobbie and Patricia Riddle[38]	2008	Gbest, pbest	Clustering
22	Juan, Jos, Nebro1 and Carlos A. Coello Coello[39]	2009	velocity constraint mechanism	-
23	L. Benameur, J.Alami, A. El Imrani[44]	2009	Pareto dominance, fuzzy clustering	Clustering
24	Bin XU Jing YU* YouGan ZHU[45]	2010	escalating strategy	--
25	Yunxia Pei[46]	2010	a grid workflow scheduling algorithm	workflow applications in Grids
26	Moayed Daneshyari, Gary G. Yen[47]	2011	cultural framework	Benchmark functions

Table 1. (continued)

27	Junwan Liu, Yiming Chen[48]	2010	Sigma method ϵ -dominance	Clustering
28	Li Zhongkai, Zhu Zhencai, Zhang Huiqin[49]	2010	crowding distance sorting, dominance Pareto	Benchmark functions
29	Hsing Hung Lin[50]	2010	Gbest, pbest Pareto set	Benchmark functions open-shop scheduling problems
30	Gary G. Yen and Wen Fung Leong[51]	2010	multiobjective constraint handling technique	Benchmark functions
31	Murilo R. Pontes, Fernando B. Lima Neto, Carmelo J. A. Bastos-Filho[53]	2011	auto-adaptation capability	Benchmark functions
32	Carmelo J. A. Bastos-Filho, P'ericles B. C. Miranda[54]	2011		
33	Jintao Yao, Bo Yang, Mingwu Zhang, Yuyan Kong [52]	2011	Predatory Escaping Behavior, information sharing, Pareto based	Benchmark functions
34	Nor ,Mohamad, Ammar[43]	2011	Penality function	-
35	Li Ransikarn Esraa [42]	2011	Attraction and Repulsion (AR) mechanism., crowding distance	-

3 Conclusion

PSO is a useful optimization algorithm. The basic PSO is proposed for optimization of single objective continuous problem . In more recent works the concept of PSO has been expanded to allow it to handle other optimization problems such as; binary, discrete, combinatorial, constrained and multiobjective optimization. This paper reviewed some of the works conducted in constrained and multiobjective optimization problems. This body of work reviewed suggest the significance of PSO as optimization strategy and highlights its evolution from being used for simple single objective continuous problems to more complex multiobjective and constrained problem

References

1. Kennedy, J., Eberhart, R.C.: Particle Swarm Optimization. In: Proceedings of the 1995 IEEE International Conference on Neural Networks, pp. 1942–1948. IEEE Service Center, Piscataway (1995)
2. Eberhart, R., Kennedy, J.: A new optimizer using particle swarm theory. In: Proc. 6th Int. Symp. Micro Machine and Human Science (MHS), pp. 39–43 (October 1995)

3. Kennedy, J., Eberhart, R.C.: A Discrete Binary Version of the Particle Swarm Algorithm. In: Proceedings of the 1997 IEEE Conference on Systems, Man, and Cybernetics, pp. 4104–4109. IEEE Service Center, Piscataway (1997)
4. Shi, Y., Eberhart, R.C.: Parameter Selection in Particle Swarm Optimization. In: Porto, V.W., Sarava-nan, N., Waagen, D., Eibe, A. (eds.) EP 1998. LNCS, vol. 1447, pp. 591–600. Springer, Heidelberg (1998)
5. Eberhart, R., Shi, Y.: Comparison between Genetic Algorithms and Particle Swarm Optimization. In: Porto, V.W., Saravanan, N., Waagen, D., Eibe, A. (eds.) EP 1998. LNCS, vol. 1447, pp. 611–619. Springer, Heidelberg (1998)
6. Moore, J., Chapman, R.: Application of Particle Swarm to Multiobjective Optimization. In: Department of Computer Science and Software Engineering, Auburn University, (unpublished manuscript) (1999)
7. Kennedy, J., Eberhart, R.C.: Swarm Intelligence. Morgan Kaufmann Publishers, San Francisco (2001)
8. Ray, T., Liew, K.: A Swarm Metaphor for Multiobjective Design Optimization. *Engineering Optimization* 34(2), 141–153 (2002)
9. Whitley, D., Goldberg, D., Cantú-Paz, E., Spector, L., Parmee, I., Beyer, H.-G.: Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2000), pp. 771–777. Morgan Kaufmann, San Francisco (2000)
10. Parsopoulos, K., Vrahatis, M.: Particle Swarm Optimization Method in Multiobjective Problems. In: SAC 2002, pp. 603–607. ACM Press (2002)
11. Jin, Y., Okabe, T., Sendhoff, B.: Dynamic Weighted Aggregation for Evolutionary Multi-Objective Optimization: Why Does It Work and How? In: Spector, L., Goodman, E.D., Wu, A., Langdon, W., Voigt, H.-M., Gen, M., Sen, S., Dorigo, M., Pezeshk, S., Garzon, M.H., Burke, E. (eds.) Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2001), pp. 1042–1049. Morgan Kaufmann Publishers, San Francisco (2001)
12. Parsopoulos, K., Tasoulis, D., Vrahatis, M.: Multiobjective Optimization Using Parallel Vector Evaluated Particle Swarm Optimization. In: Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2004), vol. 2, pp. 823–828. ACTA Press, Innsbruck (2004)
13. Schaffer, J.D.: Multiple Objective Optimization with Vector Evaluated Genetic Algorithms. In: Genetic Algorithms and their Applications: Proceedings of the First International Conference on Genetic Algorithms, pp. 93–100. Lawrence Erlbaum, Hillsdale (1985)
14. Hu, X., Eberhart, R.: Multiobjective Optimization Using Dynamic Neighborhood Particle Swarm Optimization. In: Congress on Evolutionary Computation (CEC 2002), Piscataway, New Jersey, vol. 2, pp. 1677–1681. IEEE Service Center (May 2002)
15. Hu, X., Eberhart, R.C., Shi, Y.: Particle Swarm with Extended Memory for Multiobjective Optimization. In: 2003 IEEE Swarm Intelligence Symposium Proceedings, Indianapolis, Indiana, USA, pp. 193–197. IEEE Service Center (April 2003)
16. Fieldsend, J.E., Singh, S.: A Multi-Objective Algorithm based upon Particle Swarm Optimisation, an Efficient Data Structure and Turbulence. In: Proceedings of the 2002 U.K. Workshop on Computational Intelligence, Birmingham, UK, pp. 37–44 (September 2002)
17. Laumanns, M., Thiele, L., Deb, K., Zitzler, E.: Combining Convergence and Diversity in Evolutionary Multi-objective Optimization. *Evolutionary Computation* 10(3), 263–282 (2002)

18. Toscano Pulido, G., Coello Coello, C.A.: Using Clustering Techniques to Improve the Performance of a Multi-objective Particle Swarm Optimizer. In: Deb, K., et al. (eds.) GECCO 2004, Part-I. LNCS, vol. 3102, pp. 225–237. Springer, Heidelberg (2004)
19. Coello Coello, C.A., Toscano Pulido, G., Salazar Lechuga, M.: Handling Multiple Objectives With Particle Swarm Optimization. *IEEE Transactions on Evolutionary Computation* 8(3), 256–279 (2004)
20. Baltar, A.M., Fontane, D.G.: A generalized multiobjective particle swarm optimization solver for spreadsheet models: application to water quality. In: *Hydrology Days 2006*, Fort Collins, Colorado, USA (March 2006)
21. Tayal, M.: Particle Swarm Optimization for Mechanical Design. Master's thesis, The University of Texas at Arlington, Arlington, Texas, USA (December 2003)
22. Mostaghim, S., Teich, J.: Strategies for Finding Good Local Guides in Multi-objective Particle Swarm Optimization (MOPSO). In: 2003 IEEE Swarm Intelligence Symposium Proceedings, pp. 26–33. IEEE Service Center, Indianapolis (2003)
23. Mostaghim, S., Teich, J.: Covering Pareto-optimal Fronts by Subswarms in Multi-objective Particle Swarm Optimization. In: 2004 Congress on Evolutionary Computation (CEC 2004), vol. 2, pp. 1404–1411. IEEE Service Center, Portland (2004)
24. Li, X.: A Non-dominated Sorting Particle Swarm Optimizer for Multiobjective Optimization. In: Cantú-Paz, E., Foster, J.A., Deb, K., Davis, L., Roy, R., O'Reilly, U.-M., Beyer, H.-G., Kendall, G., Wilson, S.W., Harman, M., Wegener, J., Dasgupta, D., Potter, M.A., Schultz, A., Dowsland, K.A., Jonoska, N., Miller, J., Standish, R.K. (eds.) GECCO 2003, Part-I. LNCS, vol. 2723, pp. 37–48. Springer, Heidelberg (2003)
25. Li, X.: Better Spread and Convergence: Particle Swarm Multiobjective Optimization Using the Maximin Fitness Function. In: Deb, K., et al. (eds.) GECCO 2004, Part-I. LNCS, vol. 3102, pp. 117–128. Springer, Heidelberg (2004)
26. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* 6(2), 182–197 (2002)
27. Balling, R.: The Maximin Fitness Function; Multi-objective City and Regional Planning. In: Fonseca, C.M., Fleming, P.J., Zitzler, E., Deb, K., Thiele, L. (eds.) EMO 2003. LNCS, vol. 2632, pp. 1–15. Springer, Heidelberg (2003)
28. Srinivasan, D., Seow, T.H.: Particle Swarm Inspired Evolutionary Algorithm (PS-EA) for Multiobjective Optimization Problem. In: Proceedings of the 2003 Congress on Evolutionary Computation (CEC 2003), vol. 4, pp. 2292–2297. IEEE Press, Canberra (2003)
29. Srinivasan, D., Seow, T.H.: Particle Swarm Inspired Evolutionary Algorithm (PS-EA) for Multi-Criteria Optimization Problems. In: Abraham, A., Jain, L., Goldberg, R. (eds.) Evolutionary Multiobjective Optimization: Theoretical Advances And Applications, pp. 147–165. Springer, London (2005) ISBN 1-85233-787-7
30. Coello Coello, C.A., Salazar Lechuga, M.: MOPSO: A Proposal for Multiple Objective Particle Swarm Optimization. In: Congress on Evolutionary Computation (CEC 2002), vol. 2, pp. 1051–1056. IEEE Service Center, Piscataway (2002)
31. Zhang, L., Zhou, C., Liu, X., Ma, Z., Liang, Y.: Solving Multi Objective Optimization Problems Using Particle Swarm Optimization. In: Proceedings of the 2003 Congress on Evolutionary Computation (CEC 2003), vol. 4, pp. 2400–2405. IEEE Press, Canberra (2003)
32. Baumgartner, U., Magele, C., Renhart, W.: Pareto Optimality and Particle Swarm Optimization. *IEEE Transactions on Magnetics* 40(2), 1172–1175 (2004)

33. Chow, C., Tsui, H.: Autonomous Agent Response Learning by a Multi-Species Particle Swarm Optimization. In: 2004 Congress on Evolutionary Computation (CEC 2004), vol. 1, pp. 778–785. IEEE Service Center, Portland (2004)
34. Ho, S., Yang, S., Ni, G., Lo, E.W., Wong, H.: A Particle Swarm Optimization-Based Method for Multiobjective Design Optimizations. *IEEE Transactions on Magnetics* 41(5), 1756–1759 (2005)
35. Mahfouf, M., Chen, M.-Y., Linkens, D.A.: Adaptive Weighted Particle Swarm Optimisation for Multi-objective Optimal Design of Alloy Steels. In: Yao, X., Burke, E.K., Lozano, J.A., Smith, J., Merelo-Guervós, J.J., Bullinaria, J.A., Rowe, J.E., Tiño, P., Kabán, A., Schwefel, H.-P. (eds.) PPSN 2004, Part-VIII. LNCS, vol. 3242, pp. 762–771. Springer, Heidelberg (2004)
36. Zhao, B., Cao, Y.J.: Multiple objective particle swarm optimization technique for economic load dispatch. *Journal of Zhejiang University Science* 6A(5), 420–427 (2005)
37. Dehuria, S., Chob, S.B.: Multi-criterion Pareto based particle swarm optimized polynomial neural network for classification: A review and state-of-the-art computer science review. Elsevier (2008)
38. Alam, S., Dobbie, G., Riddle, P.: An Evolutionary Particle Swarm Optimization Algorithm for Data Clustering. In: 2008 IEEE Swarm Intelligence Symposium, St. Louis MO USA, September 21-23 (2008)
39. Juan, Jos, Nebro, Coello Coello, C.A.: Multi-Objective Particle Swarm Optimizers: An Experimental Comparison (2009)
40. Guliashki, V., Toshev, H., Korsemov, C.: Survey of Evolutionary Algorithms Used in Multiobjective Optimization bulgarian academy of sciences problems of engineering cybernetics and robotics, sofia (2009)
41. Krami, N., El-Sharkawi, M.A., Akherraz, M.: Multi Objective Particle Swarm Optimization Technique for Reactive Power Planning. In: 2006 Swarm Intelligence Symposium (SIS 2006), pp. 170–174. IEEE Press, Indianapolis (2006)
42. Li, Ransikarn, Esraa: A Novel Diversity Guided Particle Swarm Multi-objective Optimization Algorithm *International Journal of Digital Content Technology and its Applications* 5(1) (January 2011)
43. Nor, Mohamad, Ammar: Particle Swarm Optimization for Constrained and Multiobjective Problems: A Brief Review. In: 2011 International Conference on Management and Artificial Intelligence IPEDR, vol. 6. IACSIT Press, Bali (2011)
44. Benameur, L., Alami, J., El Imrani, A.: A New Hybrid Particle Swarm Optimization Algorithm for Handling Multiobjective Problem Using Fuzzy Clustering Technique 2009. In: IEEE 2009 International Conference on Computational Intelligence, Modelling and Simulation (2009)
45. Xu, B., Yu, J., Zhu, Y.: Multi-Objective PSO Algorithm Based on Escalating Strategy. IEEE (2010)
46. Pei, Y.: AMOPSO Approach to Grid Workflow Scheduling Asia-Pacific Conference on Wearable Computing Systems (2010)
47. Daneshyari, M., Yen, G.G.: Cultural-Based Multiobjective Particle Swarm Optimization. *IEEE Transactions on Systems, Man, and Cybernetics—PART B: Cybernetics* 41(2) (April 2011)
48. Liu, J., Chen, Y.: Dynamic Biclustering of Microarray Data with MOPSO. In: IEEE International Conference on Granular Computing (2010)
49. Li, Z., Zhu, Z., Zhang, H.: DSMOPSO: A Distance Sorting based Multiobjective Particle Swarm Optimization Algorithm. In: Sixth International Conference on Natural Computation, ICNC 2010 (2010)

50. Lin, H.H.: A Multi-objective Particle Swarm Optimization for Openshop Scheduling Problems. In: Sixth International Conference on Natural Computation, ICNC 2010 (2010)
51. Yen, G.G., Leong, W.F.: Constraint Handling Procedure for Multiobjective Particle Swarm Optimization. IEEE (2010)
52. Yao, J., Yang, B., Zhang, M., Kong, Y.: Multiobjective Particle Swarm Optimization with Predatory Escaping Behavior (2011)
53. Pontes, M.R., Lima Neto, F.B., Carmelo, J.A.: Bastos-Filho Adaptive Clan Particle Swarm Optimization
54. Bastos-Filho, C.J.A., Miranda, P.B.C.: Multi-Objective Particle Swarm Optimization using Speciation. IEEE (2011)

Analysis of Similarity Measures with WordNet Based Text Document Clustering

Nadella Sandhya¹ and A. Govardhan²

¹ CSE Dept. Gokaraju Rangaraju Institute of Engineering & Technology,
Hyderabad, 500072, India

nadella_sandhya@yahoo.co.in

² JNTUH College of Engineering, Jagtial, 505501, India

govardhan_cse@yahoo.co.in

Abstract. Text Document Clustering aids in reorganizing the large collections of documents into a smaller number of manageable clusters. While several clustering methods and the associated similarity measures have been proposed in the past, the partition clustering algorithms are reported performing well on document clustering. Usually cosine function is used to measure the similarity between two documents in the criterion function, but it may not work well when the clusters are not well separated. Word meanings are better than word forms in terms of representing the topics of documents. Thus, here we have involved ontology into the text clustering algorithm. In this research WordNet based document representation is attempted by assigning each word a part-of-speech (POS) tag and by enriching the 'bag-of-words' data representation with synset concept which corresponds to synonym set that is introduced by WordNet. After replacing the 'bag of words' with their respective Synset IDs a variant of K-Means algorithm is used for document clustering. Then we compare the three popular similarity measures (Cosine, Pearson Correlation Coefficient and extended Jaccard) in conjunction with different types of vector space representation (Term Frequency and Term Frequency-Inverse Document Frequency) of documents.

1 Introduction

Clustering of text documents was initially used for improving the precision or recall in an Information Retrieval System [1] [2]. Currently clustering has been proposed for use in browsing a collection of documents [3] or in organizing the results returned by a search engine in response to user's query [4] or help users quickly identify and focus on the relevant set of results. However, there are several challenges that clustering techniques normally have to overcome. This work focuses on the following challenges 1.choosing an appropriate similarity measure for text clustering 2.utilizing domain knowledge in text clustering.

This paper is organized as follows. The section 2 deals with the related work in text document clustering, section 3 describes the preprocessing, POS tagging and the use of WordNet for replacing words with their Synset IDs. Section 4 describes the document representation used in the experiments. Section 5 discusses the similarity

measures and their semantics. Section 6 presents the basic K-Means and modified K-Means clustering algorithm, Section 7 explains experiment settings, evaluation approaches, results and analysis and Section 8 concludes and discusses future work.

2 Related Work

Many clustering techniques have been proposed in the literature. Clustering algorithms are mainly categorized into Hierarchical and Partitioning methods [2, 3, 4, 5]. Hierarchical clustering method works by grouping data objects into a tree of clusters [6]. These methods can further be classified into agglomerative and divisive hierarchical clustering depending on whether the hierarchical decomposition is formed in a bottom-up or top-down fashion. K-Means and its variants [7, 8, 9] are the most well-known partitioning methods [10]. Here K-Means++ a partitioned based clustering technique is used on the high dimensional sparse data representing text documents.

The most experienced lexical tool in text related studies is WordNet [15]. WordNet is one of the most widely used thesauri for English. To model the lexical knowledge of a native English speaker WordNet can be used. WordNet(2.1) contains over 155,000 terms organized in 117597 synsets. It groups nouns, verbs, adjectives and adverbs into sets of synonyms called synsets, representing the underlying lexical concepts. Synonymy is the basic semantic relation between the words in the WordNet. There exists a rich set of relations between the words and the synsets and between the synsets themselves. The synsets are organized into senses, giving thus the synonyms of each word, and also into hyponym / hypernym (i.e., Is-A), and meronym / holonym (i.e., Part-Of) relationships, providing a hierarchical tree-like structure for each term. The applications of WordNet to various IR techniques have been widely researched [11]. For example in [12] they combine the WordNet knowledge with fuzzy association rules and in [13] they extend the bisecting k-means using WordNet.

In this paper a variant of K-Means [10] partitioning clustering approach based on WordNet is used for document clustering. We also analyse the comparison of three popular similarity measures (Cosine, Pearson correlation and extended Jaccard) in conjunction with different types of vector space representation (Term frequency and term frequency and inverse document frequency) of documents.

3 Document Preprocessing Using WordNet

Initially we need to preprocess the documents. This step is imperative. The next step is to analyse the prepared data and divide it into clusters using clustering algorithm. The effectiveness of clustering is improved by adding the Part-Of-Speech Tag and making use of the WordNet for synsets.

The most important procedure in the preprocessing of documents using WordNet is to enrich the term vectors with concepts from the core ontology. WordNet covers semantic and lexical relations between word forms and word meanings. The first preprocessing step is POS tagging. POS tagging is a process of assigning correct syntactic categories to each word. Tag set and word disambiguation rules are

fundamental parts of any POS tagger. The POS tagger relies on the text structure and morphological differences to determine the appropriate part-of-speech. WordNet contains only nouns, verbs, adjectives and adverbs. Since nouns and verbs are more important in representing the content of documents and also mainly form the frequent word meaning sequences, we focus only on nouns and verbs and remove all adjectives and adverbs from the documents. For those word forms that do not have entries in WordNet, we keep them in the documents since these unidentified word forms may capture unique information about the documents. We remove the stopwords and then stemming is performed. The morphology function provided with WordNet is used for stemming as it only yields stems that are contained in the WordNet dictionary and also achieves improved results than Porter stemmer. The stemmed words are then looked up in the WordNet the lexical database to replace the words by their synset IDs. The words with the same synonyms are merged and are assigned a unique ID.

These preprocessing steps aim to improve the cluster quality [14]. These steps lead to the reduction of dimensions in the term space.

4 Document Representation

The representation of a set of documents as vectors in a common vector space is known as the vector space model. Documents in vector space can be represented using Boolean, Term Frequency and Term Frequency – Inverse Document Frequency.

In Boolean representation, if a term exists in a document, then the corresponding term value is set to one otherwise it is set to zero. Boolean representation is used when every term has equal importance and is applied when the documents are of small size.

In Term Frequency and Term Frequency Inverse Document Frequency the term weights have to be set. The term weights are set as the simple frequency counts of the terms in the documents. This reflects the intuition that terms occur frequently within a document may reflect its meaning more strongly than terms that occur less frequently and should thus have higher weights.

Each document d is considered as a vector in the term-space and represented by the term frequency (TF) vector:

$$d_{tf} = [tf_1, tf_2, \dots, tf_D] \quad (1)$$

where tf_i is the frequency of term i in the document and D is the total number of unique terms in the text database.

The tf-idf representation of the document d is:

$$d_{tf-idf} = [tf_1 \log(n / df_1), tf_2 \log(n / df_2), \dots, tf_D \log(n / df_D)] \quad (2)$$

To account for the documents of different lengths, each document vector is normalized to a unit vector (i.e., $\|d_{tf-idf}\|=1$). In the rest of this paper, we assume that this vector space model is used to represent documents during the clustering. Given a set C_j of documents and their corresponding vector representations, the centroid vector c_j is defined as:

$$c_j = \frac{1}{|C_j|} \sum_{d_i \in c_j} d_i \quad (3)$$

where each d_i is the document vector in the set C_j , and j is the number of documents in cluster C_j . It should be noted that even though each document vector d_i is of unit length, the centroid vector c_j is not necessarily of unit length.

5 Similarity Measures

Document clustering groups similar documents to form a coherent cluster. However, the definition of a pair of documents being similar or different is not always clear and normally varies with the actual problem setting.

Accurate clustering requires a precise definition of the closeness between a pair of objects, in terms of either the pair wise similarity or distance [20]. A variety of similarity or distance measures have been proposed and widely applied, such as cosine similarity, Jaccard coefficient, Euclidean distance and Pearson Correlation Coefficient. The details of different similarity measures are described below.

5.1 Cosine Similarity Measure

The most commonly used is the cosine function. For two documents d_i and d_j , the similarity between them can be calculated

$$\cos(d_i, d_j) = \frac{d_i \cdot d_j}{\|d_i\| \|d_j\|} \quad (4)$$

where d_i and d_j are m -dimensional vectors over the term set $T = \{t_1, t_2, \dots, t_m\}$. Each dimension represents a term with its weight in the document, which is non-negative. As a result the cosine similarity is non-negative and bounded between $[0, 1]$. Cosine similarity captures a scale invariant understanding of similarity and is independent of document length. When the document vectors are of unit length, the above equation is simplified to:

$$\cos(d_i, d_j) = d_i \cdot d_j \quad (5)$$

When the cosine value is 1 the two documents are identical, and 0 if there is nothing in common between them. (i.e., their document vectors are orthogonal to each other).

5.2 Jaccard Coefficient

The Jaccard coefficient, which is sometimes referred to as the Tanimoto coefficient, measures similarity as the intersection divided by the union of the objects. For text document, the Jaccard coefficient compares the sum weight of shared terms to the sum weight of terms that are present in either of the two documents but are not the common terms.

$$\text{Jaccard Coff } (d_i, d_j) = \frac{d_i \cdot d_j}{\|d_i\|^2 + \|d_j\|^2 - d_i * d_j} \tag{6}$$

$$\text{Jaccard Index } (d_i, d_j) = \frac{d_i \cap d_j}{d_i \cup d_j} \tag{7}$$

The Jaccard Coefficient ranges between [0, 1]. The Jaccard value is 1 if two documents are identical and 0 if the two documents are disjoint. The Cosine Similarity may be extended to yield Jaccard Coefficient in case of Binary attributes.

5.3 Pearson Correlation Coefficient

Correlation is a technique for investigating the relationship between two quantitative, continuous variables, for example, age and blood pressure. Pearson's correlation coefficient (r) is a measure of the strength of the association between the two variables. There are different forms of Pearson Correlation Coefficient (PCC) formula. It is given by

$$\text{Pearson Similarity } (d_i, d_j) = \frac{m \sum_k d_{ik} X d_{jk} - TF_i X TF_j}{\sqrt{[m \sum_k d_{ik}^2 - TF_i^2][m \sum_k d_{jk}^2 - TF_j^2]}} \tag{8}$$

Where $TF_i = \sum_k d_{ik}$ and $TF_j = \sum_k d_{jk}$ m is the no. of terms in document d.

The measure ranges from +1 to -1. Positive correlation indicates that both variables increase or decrease together, whereas negative correlation indicates that as one variable increases, so the other decreases, and vice versa. When Pearson Similarity is ±1 the two documents are identical and there is no relation between variables if it is equal to zero.

The Euclidean distance is a distance measure, while the cosine similarity, Jaccard coefficient and Pearson coefficient are similarity measures. We apply a simple transformation to convert the similarity measure to distance values. Because both cosine similarity and Jaccard coefficient are bounded in [0, 1] and monotonic, we take $D = 1 - SIM$ as the corresponding distance value. For Pearson coefficient, which ranges from -1 to +1, we take $D = 1 - SIM$ when $SIM \geq 0$ and $D = |SIM|$ when $SIM < 0$.

6 Clustering Algorithm

The simple Lloyd's algorithm[17] usually referred as simple k means was first developed in 1967. This is an iterative Partitional clustering process that aims to minimize the least squares error criterion [6]. The standard k-means algorithm works as follows. Given a set of data objects D and a pre-specified number of clusters k, k

data objects are randomly selected to initialize k clusters, each one being the centroid of a cluster. The remaining objects are then assigned to the cluster represented by the nearest or most similar centroid. Next, new centroids are recomputed for each cluster and in turn all documents are re-assigned based on the new centroids. This step iterates until a converged and fixed solution is reached, where all data objects remain in the same cluster after an update of centroids. The generated clustering solutions are locally optimal for the given data set and the initial seeds.

6.1 Modified K-Means

In this paper, the basic K-Means algorithm is augmented with a special initialization technique [16] that aims at improving both the accuracy and the speed of k -means. As mentioned above we will use the modified K-Means algorithm as different choices of initial seed sets can result in very different final partitions. Methods for finding good starting points have been proposed [10]. Arthur and Vassilvitskii proposed the k -means++ algorithm, which uses a randomized Seeding technique [10].

1. Choose an initial centroid c_1 uniformly at random from D .
2. Choose the next centroid c_i , selecting $c_i = d' \in D$ with probability

$$\frac{S(d')^2}{\sum_{d \in D} S(d)^2}. \quad (9)$$

Here $S(d)$ represents the max similarity value between document D and already chosen centroids.

3. Repeat Step 2 until k centroids are chosen.
4. Assign all points to the closest centroid.
5. Recompute the centroid of each cluster.
6. Repeat steps 4 and 5 until the centroids don't change.

7 Experiment

The aim of this work is to explore the benefits of partial disambiguation of words by their POS and the inclusion of WordNet concepts. There is no systematic comparative study of the impact of similarity measures on cluster quality. This may be because the popular cost criteria do not readily translate across qualitatively different measures. It is very difficult to conduct a systematic study comparing the impact of similarity measures on cluster quality with word sense disambiguation of the documents, because objectively evaluating cluster quality is difficult in itself. In practice, manually assigned category labels are usually used as baseline criteria for evaluating clusters.

7.1 Dataset

This work experiments with a bench mark dataset Classic dataset collected from uci.kdd repositories. Classic dataset consists of four different collections CACM, CISI, CRAN and MED. We have considered 800 documents of the total 7095 documents.

In this dataset, some of the documents consists single word only, so it is meaningless to take such documents for document dataset. For eliminating these invalid documents we apply file reduction on each category, which returns the documents that supports mean length of each category. For file reduction we construct the Boolean matrices of all documents category wise and calculate mean length of each category and removed the documents from the dataset which doesn't support mean length. By this we got valid documents. From these valid documents we have collected 800 documents of four categories each. From Classic dataset 200 documents of each category again totaling to 800 documents.

This work is also experimented with an Abstracts dataset that consists of abstracts from four different fields which are downloaded from the web. Here the aim is to conduct WordNet based clustering of the downloaded abstracts from different research related topics. We have collected the abstracts of the following four research topics: Network Security, Image Processing, Natural Language Processing and Data Mining. 100 documents of each research topic are selected totaling to 400 abstract documents.

Preprocessing of the documents is performed as explained in Section 3.

7.2 Evaluation

Entropy is used as a measure of quality of the clusters. Let CS be a clustering solution. For each cluster, the class distribution of the data is calculated first, i.e., for cluster j we compute p_{ij} , the "probability" that a member of cluster j belongs to class i . Then using this class distribution, the entropy of each cluster j is calculated using the standard formula

$$E_j = - \sum p_{ij} \log(p_{ij}) \quad (10)$$

where the sum is taken over all classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each cluster weighted by the size of each cluster:

$$E_{cs} = \sum_{j=1}^m \frac{n_j * E_j}{n} \quad (11)$$

where n_j is the size of cluster j , m is the number of clusters, and n is the total number of data points.

7.3 Results Analysis

The seed points are chosen using a modified k-means thus improving the efficiency. In our previous study Boolean representation with these similarity measures did not

perform better. Also Euclidean measure performs worst [21]. So, here we have analysed these clusters for Classic dataset using Term Frequency and TF-IDF representation with Cosine, Jaccard and Pearson Corelation Coefficient measures. As shown in Tables 1, 2 PCC measure performs better with both Term Frequency and TF-IDF representations. We also observe from tables 3 that Cosine Measure performs well with TF-IDF representation. From our results it is observed that the overall entropy representation table for frequency count with Cosine shows NaN values as some of the clusters are empty. On an average, the Jaccard and Pearson measures are slightly better in generating more coherent clusters, which means the clusters have lower entropy scores. Tables 4, 5 shows partitions as generated by the Frequency Count representation and Tables 6,7 shows partition as generated by the TF-IDF using Classic dataset.

Table 1. TF-IDF Entropy Results using Classic dataset

	Cosine	Jaccard	Pearson
Cluster[0]	0.224	0.049	0.152
Cluster[1]	0.016	0.248	0.089
Cluster[2]	0.084	0.0519	0.027
Cluster[3]	0.0446	0.176	0.172

Table 2. Frequency Count Entropy Results using Classic dataset

	Cosine	Jaccard	Pearson
Cluster[0]	0.339	0.116	0.127
Cluster[1]	0.263	0.269	0.0755
Cluster[2]	0.1856	0.0348	0.0268
Cluster[3]	NaN	0.2416	0.1703

Table 3. Total Entropy Results using Classic dataset

	Cosine	Jaccard	Pearson
Frequency Count	NaN	0.1926	0.106
TF-IDF	0.111	0.1496	0.117

Table 4. Clustering Results from Frequency Count representation for Jaccard Measure using Classic dataset

	CACM	CISI	CRAN	MED	Label
Cluster[0]	18	13	192	57	CRAN
Cluster[1]	173	73	5	10	CACM
Cluster[2]	4	112	0	0	CISI
Cluster[3]	5	2	3	133	MED

Table 5. Clustering Results from Frequency Count representation for PCC Measure using Classic dataset

	CACM	CISI	CRAN	MED	Label
Cluster[0]	188	97	3	3	CACM
Cluster[1]	6	3	192	5	CRAN
Cluster[2]	0	98	0	0	CISI
Cluster[3]	6	2	5	192	MED

Table 6. Clustering Results from TF-IDF representation for Jaccard Measure using Classic dataset

	CACM	CISI	CRAN	MED	Label
Cluster[0]	20	3	196	43	CRAN
Cluster[1]	175	25	2	4	CACM
Cluster[2]	3	172	0	0	CISI
Cluster[3]	2	0	2	153	MED

Table 7. Clustering Results from TF-IDF representation for PCC Measure using Classic dataset

	CACM	CISI	CRAN	MED	Label
Cluster[0]	1	160	0	1	CISI
Cluster[1]	1	2	183	5	CRAN
Cluster[2]	194	34	7	2	CACM
Cluster[3]	4	4	10	192	MED

For the Abstracts dataset as shown in Tables 8 and 9 PCC measure performs better with both Term Frequency and TF-IDF representations. It is also observed from table 10 that Jaccard measure performs well with TF-IDF representation. On an average, the Jaccard measure is slightly better in generating more coherent clusters, which means the clusters have lower entropy scores. Tables 11 and 12 shows one partition as generated by the frequency count representation for Jaccard and Pearson measures using abstracts dataset. Tables 13 and 14 shows one partition as generated by the TF-IDF representation for Jaccard and Pearson measures.

Table 8. TF-IDF Entropy Results using Abstracts dataset

	Cosine	Jaccard	Pearson
Cluster[0]	0.2320	0.2244	0.3377
Cluster[1]	0.3312	0.1283	0.0604
Cluster[2]	0.2936	0.2767	0.2067
Cluster[3]	0.3032	0.1873	0.3614

Table 9. Term Frequency Entropy Results using Abstracts dataset

	Cosine	Jaccard	Pearson
Cluster[0]	0.3256	0.2155	0.3481
Cluster[1]	0.0358	0.0669	0.0376
Cluster[2]	0.1857	0.2718	0.2021
Cluster[3]	0.3641	0.1945	0.3240

Table 10. Total Entropy Results using Abstracts dataset

	Cosine	Jaccard	Pearson
Frequency Count	0.2929	0.2131	0.2573
TF-IDF	0.2365	0.1916	0.2446

Table 11. Clustering Results from Frequency Count representation for Jaccard Measure using Abstracts dataset

	IP	DM	NLP	NS	Label
Cluster[0]	1	94	4	1	DM
Cluster[1]	1	4	3	97	NS
Cluster[2]	0	2	91	1	NLP
Cluster[3]	98	0	2	1	IP

Table 12. Clustering Results from Frequency Count representation for PCC Measure using Abstracts dataset

	IP	DM	NLP	NS	Label
Cluster[0]	1	4	1	95	NS
Cluster[1]	0	78	0	1	DM
Cluster[2]	3	18	96	4	NLP
Cluster[3]	96	0	3	0	IP

Table 13. Clustering Results from TF-IDF representation for Jaccard Measure using Abstracts dataset

	IP	DM	NLP	NS	Label
Cluster[0]	1	91	6	2	DM
Cluster[1]	1	3	3	96	NS
Cluster[2]	1	6	90	2	NLP
Cluster[3]	97	0	1	0	IP

Table 14. Clustering Results from TF-IDF representation for PCC Measure using Abstracts dataset

	IP	DM	NLP	NS	Label
Cluster[0]	1	3	2	96	NS
Cluster[1]	0	80	0	1	DM
Cluster[2]	2	17	95	3	NLP
Cluster[3]	97	0	3	0	IP

The clustering accuracy is used as a measure of a clustering result. Clustering accuracy r is defined as

$$r = \frac{\sum_{i=1}^4 a_i}{n} \quad (12)$$

where a_i is the number of instances occurring in both cluster i and its corresponding class and n is the number of instances in the dataset. The clustering accuracy is more for TF-IDF representation with Pearson's and Jaccard coefficient measures. The Classic dataset has shown above 91 percent accuracy for TF-IDF representation with Pearson measure. For Abstracts dataset the cluster accuracy is above 93 percent.

8 Conclusions and Future Work

We have enhanced the document clustering with background knowledge from an external database, WordNet by applying semantics knowledge to the document representations to represent relationships between the terms and studied the effect of the three similarity measures exhaustively.

In this study we found that all the measures have significant effect on Partitional clustering of text documents. Pearson correlation coefficient is slightly better as the resulting clustering solutions are more balanced and is nearer to the manually created categories. Here the experiments using WordNet for document clustering always resulted in consistent cluster accuracy, of above 90% with TF-IDF representation with Pearson measure. The Jaccard and Pearson coefficient measures find more coherent clusters. Finally the TF-IDF representation shows better results with both Jaccard and Pearson measure. Considering the type of cluster analysis involved in this study, we can see that there are four components that affect the final results—representation of the documents, distance or similarity measures considered, using background knowledge for enriching ‘bag of words’ representation and the clustering algorithm itself. In our future work we would like to explore WordNet more in order to apply it with text document clustering.

References

1. Van Rijsbergen, C.J.: *Information Retrieval*, 2nd edn. Butterworth, London (1989)
2. Kowalski, G.: *Information Retrieval Systems – Theory and Implementation*. Kluwer Academic Publishers (1997)
3. Cutting, D.R., Karger, D.R., Pedersen, J.O., Tukey, J.W.: Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. In: SIGIR 1992, pp. 318–329 (1992)
4. Zamir, O., Etzioni, O., Madani, O., Karp, R.M.: Fast and Intuitive Clustering of Web Documents. In: KDD 1997, pp. 287–290 (1997)
5. Koller, D., Sahami, M.: Hierarchically classifying documents using very few words. In: Proceedings of the 14th International Conference on Machine Learning (ML), pp. 170–178 (1997)
6. Salton, G.: *Automatic Text Processing*. Addison-Wesley, New York (1989)
7. Steinbach, M., Karypis, G., Kumar, V.: A Comparison of Document Clustering Techniques. In: KDD Workshop on Text Mining (2000)
8. Cutting, D.R., Pedersen, J.O., Karger, D.R., Tukey, J.W.: Scatter/gather: A cluster-based approach to browsing large document collections. In: Proceedings of the ACM SIGIR (1992)
9. Larsen, B., Aone, C.: Fast and Effective Text Mining using Linear-time Document Clustering. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (1999)
10. Arthur, D., Vassilvitskii, S.: K-means++ the advantages of careful seeding. In: Symposium on Discrete Algorithms (2007)
11. Varelas, G., Voutsakis, E., Raftopoulou, P., Petrakis, E., Milios, E.: Semantic similarity methods in wordNet and their application to information retrieval on the web. In: Workshop On Web Information And Data Management, Proceedings of the 7th annual ACM international workshop on Web information and data management, pp. 10–16 (2005)
12. Chen, C.-L., Tseng, F.S.C., Liang, T.: An Integration of Fuzzy Association Rules and WordNet for Document Clustering. In: Theeramunkong, T., Kijirikul, B., Cercone, N., Ho, T.-B. (eds.) PAKDD 2009. LNCS, vol. 5476, pp. 147–159. Springer, Heidelberg (2009)
13. Sedding, J., Kazakov, D.: WordNet-based text document clustering. In: Proc. of COLING-Workshop on Robust Methods in Analysis of Natural Language Data (2004)
14. Sedding, J., Kazakov, D.: WordNet-based Text Document Clustering
15. Miller, G.: WordNet: a lexical database for English. *Communications of the ACM* 38(11), 39–41 (1995)
16. Technische Universität Dresden, An Empirical Study of K-Means Initialization Methods for Document Clustering
17. Lloyd, S.P.: Least squares quantization in pcm. *IEEE Transactions on Information Theory* 28(2), 129–137 (1982)
18. Strehl, A., Ghosh, J., Mooney, R.: Impact of similarity measures on web-page clustering. In: AAAI-2000: Workshop on Artificial Intelligence for Web Search (July 2000)
19. Huang, A.: Similarity Measures for Text Document Clustering. In: The Proceedings of New Zealand Computer Science Research Student Conference (2008)
20. Sandhya, N., Srilalitha, Y., Anuradha, K., Govardhan, A.: Analysis of stemming algorithm for Text Clustering

Application of Particle Swarm Optimization for Combined Environmental and Economic Dispatch of IEEE 30 Bus System Using Fuzzy Logic Technique

Sankaramurthy Padmini, Teresa George, and Medepalli Sandeep

Department of Electrical and Electronics Engineering,
SRM University, Chennai, India
{padminisp81, treesag77, sandeep_med}@gmail.com

Abstract. In this paper a method has been proposed to solve multi-objective optimization method using fuzzy decision satisfaction method while the objectives are minimized individually using Particle Swarm Optimization. The fossil fuel plants pollutes environment by emitting some toxic gases. But this load allocation may lead to increase in the operating cost of the generating units. So, it is necessary to find out a solution which gives a balanced result between emission and cost. Thus the objective of reactive power optimization problem can be seen as minimization of real power loss over the transmission lines. All these objectives are to be met for efficient operation and control. In this project an attempt has been made to optimize each objective individually using Particle Swarm Optimization. Hence an algorithm has been developed for optimization of each objective and is then tested on IEEE 30 system. Simulation results of IEEE 30 bus network are presented to show the effectiveness of the proposed method. The results clearly show that the proposed method gives global optimum solution compared to the other methods.

Keywords: Economic dispatch, Emission dispatch, Reactive power, Voltage stability, PSO.

1 Introduction

Power system should be operated in such a fashion that simultaneously real and reactive power is optimized. Real power optimization problem is the traditional economic dispatch which minimizes the real power generation cost. Reactive power should be optimized to provide better voltage profile as well as to reduce total system transmission loss. Thus the objective of reactive power optimization problem can be seen as minimization of real power loss over the transmission lines. Traditional Economic Dispatch [1] aims at scheduling committed generating unit's outputs to meet the load demand at minimum fuel cost while satisfying equality and inequality constraints. On the other hand thermal power plants create environmental pollution by emitting toxic gases such as carbon dioxide (CO₂), sulphur dioxide (SO₂), nitrogen oxides (NO_x). Several strategies for minimizing these emissions have been proposed among which dispatch of generating units to minimize emissions as well as fuel cost is the most attractive approach as this can be applied to the traditional economic

dispatch algorithm with slight modification[5]. The four objectives of minimization of fuel cost, minimization of emission, minimization of losses and minimization of system stability index are conflicting and non commensurable. Hence trade off solution using fuzzy min-max approach is proposed in this thesis. Assuming the decision maker (DM) has imprecise or fuzzy goals of satisfying each of the objectives, the multi-objective problem can be formulated as a fuzzy satisfaction maximization problem which is basically a min-max problem [10].

2 Problem Formulation

2.1 Economic Dispatch

The ED problem is to determine the optimal combination of power outputs of all generating units to minimize the total fuel cost while satisfying the load demand and operational constraints.

$$\text{Minimize } F_T = \sum_{i=1}^n F_i (P_i) \tag{1}$$

Where F_T = Total cost of generation (Rs/hr)

n = Number of generators

P_i = Real power generation of i^{th} generator

f_i = Fuel cost function of i^{th} generator.

$$F_T = \sum_{i=1}^n F_i(P_i) = \sum_{i=1}^n a_i + b_i P_i + c_i P_i^2 \tag{2}$$

Where a_i , b_i and c_i are fuel cost coefficients

A. Equality Constraint

Equilibrium is only met when the total system generation ($\sum P_i$) equals to the total system load (PD) plus the system losses (P_{Loss})

$$\sum_{i=1}^n P_i = P_D + P_L \tag{3}$$

Where, P_D : total system demand (MW)

P_{loss} : transmission loss of the system (MW)

B. Network Losses

In the B coefficients method, network losses are expressed as a quadratic function:

$$P_L = \sum_{i=1}^n \sum_{j=1}^n P_i B_{ij} P_j + \sum_i P_i B_{i0} + B_{00} \tag{4}$$

Where, B_{ij} are constants called B coefficients or loss coefficients.

C. Inequality Constraint

These constraints reduce our permissible generator operating region to within two bounds.

$$P_{i\min} \leq P_i \leq P_{i\max}$$

Where, $P_{i,\min}$: minimum power output limit of i^{th} generator (MW)

$P_{i,\max}$: maximum power output limit of i^{th} generator (MW)

PF_i is the penalty factor of unit i given by $PF_i = \frac{1}{1 - \partial P_L / \partial P_i}$, and $\partial P_L / \partial P_i$ is the incremental loss of unit i .

power output of i^{th} unit is given as

$$P_i = \frac{1 - \frac{a_i}{\lambda} - \sum_{j=1}^n 2B_{ij}P_j}{\frac{2b_i}{\lambda} + 2B_{ii}} \tag{5}$$

D. Evaluation Function

In order to emphasize the “best” chromosome and speed up convergence of the iteration procedure, the evaluation value is normalized into the range between 0 and 1.

$$f = \frac{1}{1 + k \left(\frac{\sum_{i=1}^n P_i - P_D - P_{loss}}{P_D} \right)} \tag{6}$$

where, k is a scaling constant ($k = 50$ in this study).

2.2 Emission Dispatch

The emission dispatch problem can be defined as the following optimization problem [4]

$$\text{Minimize } E = \sum_{i=1}^n \alpha_i + \beta_i P_i + \gamma_i P_i^2 \tag{7}$$

Where

E : total emission release (Kg/hr)

$\alpha_i, \beta_i, \gamma_i$: emission coefficients of the i^{th} generating unit

Subject to demand constraint (6) and generating capacity limits (7).

$$P_{imin} \leq P_i \leq P_{imax} \tag{8}$$

The well know solution method to this problem using the coordination equation is

$$PF_i \frac{dF_i(P_i)}{dP_i} = \dots\dots\dots = PF_n \frac{dF_n(P_n)}{dP_n} \tag{9}$$

Where $\frac{dF_i(P_i)}{dP_i}$ is the incremental cost denoted by $\lambda = b_i + 2c_i$ (10)

PF_i is the penalty factor of unit i given by $PF_i = \frac{1}{1 - \partial P_L / \partial P_i}$,

and $\partial P_L / \partial P_i$ is the incremental loss of unit i . From Eq. (9) and (10) power output of i^{th} unit is given as

$$P_i = \frac{1 - \frac{\alpha_i}{\lambda_{emission}} - \sum_{j=1}^n 2B_{ij}P_j}{\frac{2\beta_i}{\lambda_{emission}} + 2B_{ii}} \tag{11}$$

Network losses are expressed as a quadratic function:

$$P_L = \sum_{i=1}^n \sum_{i=1}^n P_i B_{ij} P_j + \sum_i P_i B_{i0} + B_{00} \tag{12}$$

Where, B_{ij}, B_{i0}, B_{00} are constants called **B** coefficients or loss coefficients.

2.3 Reactive Power

The objective of RPD is to identify the reactive power control variables, which minimizes the Real power loss (P_{loss}) of the system [6]

Minimize $F = [f_1]$

$$f_1 = P_{loss} = \sum_{\substack{m \neq n \\ m, n \in S}} g_k (V_m^2 + V_n^2 - 2V_m V_n \cos \theta_{mn}) \tag{13}$$

The reactive power optimization problem is subjected to the following constraints.

A. *Equality Constraints*: These constraints represent load flow equation such as

$$\begin{aligned}
 P - V_i \sum_{j=1}^{N_b} V_j (G_{ij} \cos \theta_{ij} + B_{ij} \sin \theta_{ij}) &= 0, i \in N_g - 1 \\
 Q - V_i \sum_{j=1}^{N_b} V_j (G_{ij} \sin \theta_{ij} - B_{ij} \cos \theta_{ij}) &= 0, i \in N_{LD}
 \end{aligned}
 \tag{14} \&(15)$$

B. *Inequality Constraints*

These constraints are formulated as

(i) Voltage limits - $V_i^{\min} \leq V_i \leq V_i^{\max}; i \in N_B$ (16)

(ii) Generator reactive power capability limit

$$Q_{gi}^{\min} \leq Q_{gi} \leq Q_{gi}^{\max}; i \in N_g \tag{17}$$

(iii) Capacitor reactive power generation limit

$$Q_{ci}^{\min} \leq Q_{ci} \leq Q_{ci}^{\max}; i \in N_c \tag{18}$$

(iv) Transformer tap setting limit

$$t_k^{\min} \leq t_k \leq t_k^{\max}; i \in N_T \tag{19}$$

(v) Transmission line flow limit

$$S_l \leq S_l^{\max}; l \in N_l \tag{20}$$

2.4 Voltage Stability

The L index is a quantitative measure for the estimation of the distance of the actual state of the system to the stability limit. The L index describes the stability of the complete system and is given by [11]

$$\text{Lindex}(i) = 1 - \left| \frac{\sum_{i \in n_L} \text{FLG}(j\text{-no_units}, i) * E(i)}{E(j)} \right|
 \tag{21}$$

Here n_g = number of generators
 n = no of buses.

The L index value varies in a range between 0 (no load) and 1 (voltage collapse).

3 Economic and Emission Dispatch Using PSO

3.1 Algorithm

1. Specify the lower and upper bound generation power of each unit, and calculate λ_{\max} and λ_{\min} . Initialize randomly the individuals of the population according to the limit of each unit including individual dimensions, searching points, and velocities. These initial individuals must be feasible candidate solutions that satisfy the practical operation constraints.
2. Set iteration count=1.
3. Set population count=1.
4. To each individual in the population (i.e at each λ) compute power output of all generators using Eq.(5). Employ the B-coefficient loss formula Eq.(3) to calculate the transmission loss P_L .
5. Calculate the evaluation value of each individual in the population using Eq.(6). Compare each individual's evaluation value with its P_{best} . If the evaluation value of each individual is better than the previous P_{best} , the current value is set to be P_{best} .
6. Increment individual count by 1.If count < population size go to step (4).
7. The best evaluation value among the P_{bests} is denoted as g_{best} .
8. Modify the member velocity V of each individual according to

$$v_i^{k+1} = k * (w * v_i^k + c_1 * rand_1 * (pbest_i - x_i) + c_2 * rand_2 * (gbest_i - x_i))$$

$$x_i^{k+1} = x_i + v_i^{k+1}$$
9. If $v_i^{k+1} > Vmax$, then $v_i^{k+1} = Vmax$ and if $v_i^{k+1} < -Vmax$, then $v_i^{k+1} = -Vmax$.
10. Modify the member position of each individual P_i according to

$$P_i^{(k+1)} = P_i^{(k)} + V_i^{(k+1)}$$

$$P_i^{(k+1)}$$
 must satisfy the constraints.
11. Increment iteration count by 1.If the number of iterations reaches the maximum,then go to Step 13.Otherwise, go to Step 3.
12. The individual that generates the latest g_{best} is the optimal generation power of each unit with the minimum total generation cost.
13. At this power generation compute emission release. Run FDC load flow to determine system losses and stability index.

4 Fuzzified PSO for Multiobjective Problem

Each particle consists of power generations of all units excluding slack bus voltages, taps and shunts encoded in it. The size of each particle is equal to sum of active power generations, no of voltages excluding slack bus, number of voltage, taps, and shunts. Assuming the decision maker (DM) has imprecise or fuzzy goals of satisfying each of the objectives, the multi-objective problem can be formulated as a fuzzy satisfaction maximization problem which is basically a min-max problem. [10]

4.1 Proposed Algorithm

The proposed solution strategy for the multi objective problem is shown in the following algorithm

1. Read the system data.
2. Read the values of fixed cost, loss, index, emission for each sub problems.
3. Form Y_{bus} matrix and FLG matrix for L index calculation.
4. Form B1 sub matrix. Decompose B1 by Cholesky decomposition

previous P_{best} . The current value is set to be P_{best} . If the best P_{best} is better than g_{best} , the value is set to be g_{best} .

Step 7: If the stopping criterion is reached, then print the result and stop; otherwise repeat steps 2–6.

5 Case Studies and Results for IEEE 30 Bus System

The line data, bus data cost and emission coefficients of IEEE 30 bus system. 25 independent runs are made for each sub problem and the values of four factors considered at minimum value of each sub problem over 25 independent runs are determined. These values for all sub problems are given in Table 1.

Optimization Problem	Fuel Cost (\$/hr)	Losses (MW)	Stability Index	Emission (kg/hr)
Fuel cost minimization	806.498025	10.5826	0.272567	380.671279
Losses minimization	945.214690	4.32680	0.272342	232.701959
Stability Index minimization	897.142571	33.557655	0.162446	375.611008
Emission minimization	932.094511	4.404039	0.267070	229.144834

6 Conclusions

In this work an approach to solve multi objective problem which aims at minimizing fuel cost, real power loss, emission release and improving stability index of the system simultaneously has been proposed. Several system constraints are taken care off.

We have successfully implemented Particle Swarm Optimization solution for Economic Dispatch Problem. The so algorithm has been tested on IEEE 30 bus system. An attempt has been made to determine the optimum dispatch of generators, when emission release is taken as objective. The algorithm has been tested on IEEE

30 bus system. Reactive power optimization is taken as another objective and the algorithm has been developed for minimizing the total system losses using PSO. Improving stability index of the system is taken as another independent objective and this improvement is done using PSO. Thus all the four objectives are solved individually and the results from these individual optimizations are fuzzified and final trade off solution is thus obtained. Our proposed approach satisfactorily finds global optimal solution within a small number of iterations.

References

1. AlRashidi, M.R., El-Hawary, M.E.: A Survey of Particle Swarm Optimization Applications in Electric Power Systems. *IEEE Trans. On Evolutionary Computation* (2006)
2. Park, J.-B., Lee, K.-S., Shin, J.-R., Lee, K.Y.: A Particle Swarm Optimization for Economic Dispatch with Non-smooth Cost Functions. *IEEE Trans. on Power Syst.* 20(1), 34–42 (2005)
3. Gaing, Z.-L.: Particle Swarm Optimization to Solving the Economic Dispatch Considering the Generator Constraints. *IEEE Tans. on Power Syst.* 18(3), 1187–1195 (2003)
4. Kumar, I.S., Dhanushkodi, K., Jaya Kumar, J., Kumar Charlie Paul, C.: Particle Swarm Optimization Solution to Emission and Economic Dispatch Problem. In: *IEEE TENCON* (2003)
5. Thakur, T., Sem, K., Saini, S., Sharma, S.: A Particle Swarm Optimization Solution to NO₂ and SO₂ Emissions for Environmentally Constrained Economic Dispatch Problem. In: *2006 IEEE PES Transmission and Distribution Conference and Exposition, Latin America, Venezuela* (2006)
6. Das, B., Patvardhan, C.: A New Hybrid Evolutionary Strategy for Reactive Power Dispatch. *Electric Power Research* 65, 83–90 (2003)
7. Dutta, P., Sinha, A.K.: Environmental Economic Dispatch constrained by voltage stability using PSO. In: *Electrical Engineering, IIT Kharagpur-IEEE* (2006)
8. Zhang, W., Liu, Y.: Reactive Power Optimization Based on PSO in a Practical Power System
9. Abido, M.A.: A novel multiobjective evolutionary algorithm for environmental economic power dispatch. *Electric Power Syst. Res.* 65, 71–81 (2003)
10. Rajesekaran, S., Vijayalakshmi Pai, G.A.: *Neural Networks, Fuzzy Logic and Genetic Synthesis and Applications*. Prentice Hall Pvt. Ltd. (2003)
11. Reis, C., Maciel Barbosa, F.P.: A Comparison of Voltage Stability Indices. In: *IEEE MELECON 2006, Benalmádena (Málaga), Spain, May 16-19* (2006)

An Optimal Design to Schedule the Hydro Power Generation Using Lagrangian Relaxation Method

Santhoshkumar Maheswari and C. Vijayalakshmi Seshathri

Department of Mathematics,
Sathyabama University, Chennai – 119
rohithkumar2007@gmail.com, vijusesha2002@yahoo.co.in

Abstract. This paper mainly deals with hydroelectric generation scheduling. It is formulated as a Mixed Integer Programming (MIP) model with respect to various constraints, electric power balance, water flows, hydro discharge limits and reservoir limits and then decomposed by Lagrangian Relaxation (LR) method. This method designs the schedule for the generating units which minimizes the generation cost. The objective of this scheduling algorithm is to obtain the schedule which gives the optimal amount of generated powers for the hydro units. Based on numerical calculations and graphical representations the optimal schedule can be obtained which reduces the production cost, increases the system reliability and maximizes energy capability of reservoirs.

Keywords: Generation scheduling, Mixed Integer Programming, Lagrangian Relaxation Method.

1 Introduction

In day to day life hydrothermal scheduling is a important activity for electric utilities to meet the future demand. In this paper scheduling of hydro units determine the commitment and generation of power resources over a planning horizon. Because all hydro systems are different no two hydroelectric systems in the world are alike. The reason is between the natural and man made storage on the operation of hydroelectric systems.

The co-ordination and operation of hydro plants involves, the scheduling of water releases. The long range hydro scheduling problem involves the long range forecasting of water availability and the scheduling of reservoir capacities. In this paper the scheduling period is one year.

Generation scheduling minimizes the total generation cost. It is formulated as a MIP model and it can be solved by using Lagrangian relaxation method.

Many methods have been developed to solve hydro subproblems like dynamic programming (DP), network flow, and standard mixed integer programming (MIP) methods. Dynamic programming method was discussed by A.I. Cohen et al. (1985) and W.J. Trott et al. (1973). But DP suffers for practical applications due to “dimensionality”. Network flow method is used by H. Bramhund et al (1986), R.E. Rosental (1981), C.Li.P. Jap et al. (1993), J.L. Kennington et al. (1980) and H. Habibollahzadeh et al. (1991). Disadvantage of this method, it could not deal discontinuous

operating regions and discrete operating states. Recently, combination of DP and network flow method are used in G.Li.E. Hsu et al. (1997).

Lagrangian framework is a successful method as discussed by A. Cohen et al. (1987), J.J. Shaw et al. (1985), L.A.F.M. Fesseira et al. (1989), A. Renaud (1993), S. Maheswari et al. (2010, 2011) and S.J. Wang et al. (1995).

In this paper, the solution of MIP model is obtained by Lagrangian Relaxation method. The sub problem objective includes the fixed cost and variable cost which minimize the total generation cost for hydro electric generation scheduling.

2 Mathematical Formulation

Mathematical formulation is based on the generation scheduling with respect to hydro units.

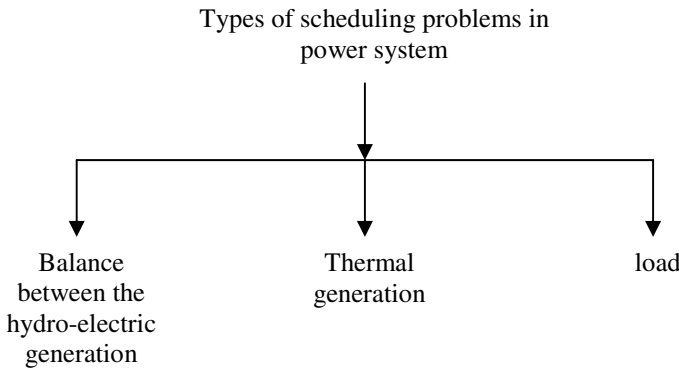


Fig. 1. Generation Scheduling

Economic scheduling is a kind of scheduling model in which water releases to satisfy all the hydro system constraints and to meet the demand for the generation of electrical energy.

3 Model Formulation

The optimization model for hydro generation scheduling can be formulated as an Mixed Integer Programming (MIP) model with respect to various constraints.

3.1 Parameters

- H – Number of hydro units
- NH – Total number of hydro units
- S – Number of pumped-storage units

PS	– Total number of pumped-storage units
t	– Time index
T	– Time horizon, $t = 1, \dots, T$
$P_d(t)$	– Power demand at time t in MW
$P_r(t)$	– System spinning reserve requirement at time t in MW
$P_{PS}(t)$	– Power generate (or) for pumping by pumped storage units PS at time t in MW
$r_H(t)$	– Spinning reserve contribution of hydro units H at time t in MW
$r_{PS}(t)$	– Spinning reserve contribution of pumped storage units PS at time t in MW
$M_H(P)$	– Maintenance cost for generation of hydro units
$A_{GH}(P)$	– The cost which is associated with the administrative and general expenses of power generation of hydro units
L_H	– Labour cost for generation of hydro units
D_{CH}	– Depreciation cost
IF_H	– The cost is related to the interest and finance charges
$SC_H(t)$	– Start up cost of hydro units H at time t
PG_H^{\min}, PG_H^{\max}	– The minimum and maximum power can be generated by the hydro units H (MW)
$R_H(t)$	– Reservoir level of hydro reservoir j at time t
R_H^{\max}	– Maximum reservoir level of hydro unit H
$u_H(t)$	– $\begin{cases} 1 & \text{if the unit H is up at time t} \\ -1 & \text{if the unit H is down at time t} \end{cases}$
R_H^{\min}	– Minimum reservoir level of hydro unit H
R_H^0	– Initial reservoir level if hydro unit H
R_H^T	– Terminal reservoir level if hydro unit H
$WD_H(t)$	– Water discharge of hydro unit H at time t
$WD_H^{\max}(t)$	– Maximum water discharge for hydro unit H at time t
$WD_H^{\min}(t)$	– Minimum water discharge for hydro unit H at time t
$X_H(t)$	– State if hydro unit H at time t, denoting number of home that the unit has been on positive (or) negative
up_H	– Minimum up time if hydro unit H, in hours
dn_H	– Minimum down time if hydro unit H, in hours

3.2 Decision Variables

$PG_H(t)$	– Amount of power generated by hydro unit H at time t
$X_H(t)$	– State if hydro unit H at time t, denoting number of home that the unit has been on positive (or) negative

3.3 Objective Function

$$MIP = \min \left\{ \sum_{t=1}^T \sum_{H=1}^{NH} [IF_H + M_H + AG_H + L_H + DC_H] \times PG_H(t) + \sum_{t=1}^T SC_H \times PG_H(t) \right\}$$

subject to

$$\sum_{H=1}^{NH} PG_H(t) + \sum_{PS=1}^{PS} P_{PS}(t) = P_d(t), \quad t = 1, \dots, T. \tag{a}$$

(system demand)

$$\sum_{t=1}^{T^{\max}} PG_H(t) \leq \sum_{t=1}^{T^{\max}} P_d(t) \tag{b}$$

$$\sum_{H=1}^{NH} r_H(t) \times PG_H(t) + \sum_{S=1}^{PS} r_{PS}(t) \times P_{PS}(t) \geq P_r(t), \quad t = 1, \dots, T \tag{c}$$

(spinning reserve requirement)

$$R_H^{\min} \leq R_H(t) \leq R_H^{\max} \tag{d}$$

(reservoir level limits)

$$R_H(t) = \begin{cases} R_H(0) = R_H^0 \\ R_H(T) = R_H^T \end{cases} \tag{e}$$

(initial & terminal reservoir levels)

$$WP_H^{\min}(t) \leq WD_H(t) \leq WD_H^{\max}(t) \tag{f}$$

$u_H(t) = 1$, minimum time of hydro unit H , $1 \leq X_H(t) \leq up_H$

$u_H(t) = -1$, minimum down time of hydro unit H , $dn_H \leq X_H(t) \leq -1$

$$\sum_{H=1}^{NH} (IF_H + AG_H + L_H) \times PG_H(t) = P_d(t) \tag{g}$$

$$\sum_{H=1}^{NH} (M_H + DC_H + L_H) \times PG_H(t) = P_d(t) \tag{h}$$

$$PG_H^{\min} \leq PG_H(t) \leq PG_H^{\max}, \quad t = 1, 2, \dots, T \tag{i}$$

4 Solution Methodology

4.1 Langrangian Relaxation Method

Relaxing the equation (a),

$$L[PG_H, P_{PS}, \lambda] =$$

$$\min \left\{ \left\{ \sum_{t=1}^T \sum_{H=1}^{NH} [IF_H + M_H + AG_H + L_H + DC_H] \times PG_H(t) + \sum_{t=1}^T SC_H \times PG_H(t) \right\} - \lambda \left\{ \sum_{H=1}^{NH} PG_H(t) + \sum_{PS=1}^{PS} P_{PS}(t) - P_d(t), \right\} \right\}$$

subject to

$$\sum_{t=1}^{T^{\max}} PG_H(t) \leq \sum_{t=1}^{T^{\max}} P_d(t) \tag{a}$$

$$\sum_{H=1}^{NH} r_H(t) \times PG_H(t) + \sum_{S=1}^{PS} r_{PS}(t) \times P_{PS}(t) \geq P_r(t), \quad t = 1, \dots, T \tag{b}$$

(spinning reserve requirement)

$$R_H^{\min} \leq R_H(t) \leq R_H^{\max} \tag{c}$$

(reservoir level limits)

$$R_H(t) = \begin{cases} R_H(0) = R_H^0 \\ R_H(T) = R_H^T \end{cases} \tag{d}$$

(initial & terminal reservoir levels)

$$WP_H^{\min}(t) \leq WD_H(t) \leq WD_H^{\max}(t) \tag{e}$$

$u_H(t) = 1$, minimum up time of hydro unit H , $1 \leq X_H(t) \leq up_H$

$u_H(t) = -1$, minimum down time of hydro unit H , $dn_H \leq X_H(t) \leq -1$

$$\sum_{H=1}^{NH} (IF_H + AG_H + L_H) \times PG_H(t) = P_d(t) \tag{f}$$

$$\sum_{H=1}^{NH} (M_H + DC_H + L_H) \times PG_H(t) = P_d(t) \tag{g}$$

$$PG_H^{\min} \leq PG_H(t) \leq PG_H^{\max}, \quad t = 1, 2, \dots, T \tag{h}$$

Lagrangian Relaxation replaces the original problem with an associated Lagrangian problem whose optimal solution provides a bound on the objective function of the problem. This is achieved by eliminating (relaxing one or more) constraints of the original model and adding these constraints, multiplied by an associated Lagrangian multiplier in the objective function.

The main objective of this method is to relax the constraints that will result in a relaxed problem. When it gives the values of multipliers, it is much easier to solve optimally. The role of these multipliers is to derive the Lagrangian problem towards a solution that satisfies the relaxed constraints.

The Lagrangian relaxation approach replaces the problem of identifying the optimal values of all the decision variables with one of finding optimal or good values for the Lagrangian multipliers. Most Lagrangian-based heuristics use a search heuristic to identify the optimal multipliers. A major benefit of Lagrangian-based heuristics is that they generate bounds (i.e., lower bounds on minimization problems and upper bounds on maximization problems) on the value of the optimal solution of the original problem.

In this paper ,Lagrangian function incudes fixed cost and variable cost which is obtained by relaxing the demand constraint from the MIP which minimizes the generation cost with respect to various constraints.

5 Numerical Calculations and Graphical Representations

Generation scheduling algorithm gives the schedule for generation of units and this implemented in MALTAB 7.0. The testing data sets are summarized in the following table.

Table 1. Optimum Generation Cost

	Kundah	Kadamparai	Erode	Tirunelveli	Total Cost
January	923.780	384.640	56.340	490.776	1855.546
February	843.633	395.043	573.581	606.645	2418.902
March	1046.688	497.412	347.113	639.474	2530.687
April	409.230	366.600	0.000	395.102	1170.942
May	661.540	225.770	0.000	368.472	1255.792
June	365.020	253.990	618.912	358.385	1596.317
July	364.150	1025.110	765.739	382.473	2537.482
August	258.430	221.560	629.000	339.403	1448.403
September	437.420	237.060	497.950	349.617	1522.054
October	330.881	245.440	652.580	440.615	1669.526
November	341.170	239.260	495.974	467.812	1544.226
December	307.070	236.370	720.020	375.300	1638.770

Table 1 gives the minimum generation cost of Rs. 1170.942 (in lakhs) in the month of April by the Scheduling algorithm. The maximum power can be utilized in April with respect the power cycle.

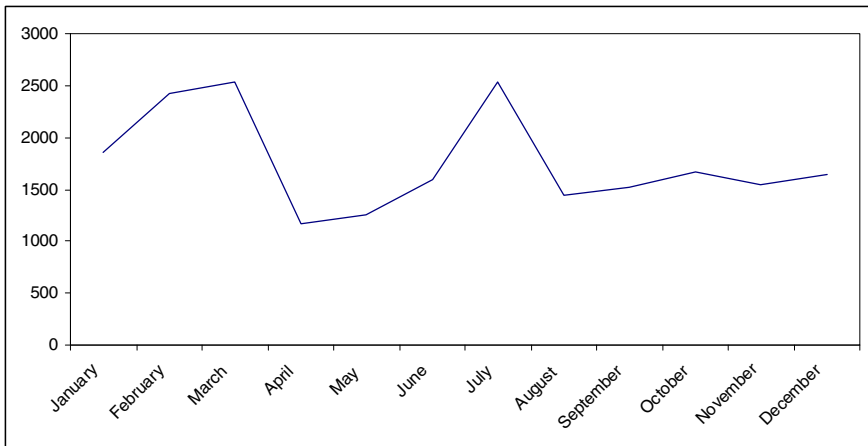


Fig. 2. Optimization Graph

6 Conclusion

In this paper, Mixed Integer Programming (MIP) model is designed with respect to various constraints. The solution is obtained by Lagrangian relaxation method. This method relaxing the constraints with respect to demand, the subproblem gives the schedule for the generation of hydro units. Minimum generation cost obtained by the schedule that is maximum power utilized in the planning period. This generation scheduling algorithm gives convergence within an acceptable execution time and highly optimal solution can be achieved.

Acknowledgement. The authors would like to thank the anonymous reviewers for their helpful suggestions and comments during the research for this paper. Authors would also like to thank Er. R.Santhosh Kumar, A.E, TNEB, Chennai, India, for his valuable support and guidance during the research for this paper.

References

1. Cohen, I., Wan, S.H.: An Algorithm for Scheduling a Large Pumped Storage Plant. *IEEE Transaction on Power Apparatus and Systems* PAS 104(8), 2099–2104 (1985)
2. Trott, W.J., Yeh, W.: Optimization of Multiple Reservoir Systems. *Journal of Hydraulics Division, ASCE* (October 1973)
3. Brannlund, H., Bubenko, J.A., Sjelvgren, D., Anderson, N.: Optimal Short Term Operation Planning of a Large Hydro-Thermal Power System Based On a Nonlinear Network Flow Concept. *IEEE Transactions on Power Systems* 1(4), 75–82 (1986)
4. Rosenthal, R.E.: A Nonlinear Network Flow Algorithm for Maximization of Benefits in a Hydroelectric Power System. *Operation Research* 29(4), 763–786 (1981)
5. Li, P.J., Streiffert, D.: Implementation of Network Flow Programming to the Hydrothermal Coordination in an Energy Management System. *IEEE Transactions on Power Systems* 8(3), 1045–1053 (1993)
6. Kennington, J.L., Helgason, R.V.: *Algorithms for Network Programming*. John Wiley & Son (1980)
7. Habibollahzadeh, H., Frances, D., Sui, U.: A New Generation Scheduling Problem at Ontario Hydro. *IEEE Transactions on Power Systems* 5(1), 65–73 (1991)
8. Li, C., Hsu, E., Svoboda, A., Tseng, C., Johnson, R.: Hydro Unit Commitment in Hydro-Thermal Optimization. In: 1996 IEEE/PES Summer Meeting, Denver, CO (July 1997), 97 SM 497-8
9. Cohen, A., Sherkat, V.: Optimization-Based Methods for Operations Scheduling. *Proceedings of IEEE* 75(12), 1574–1591 (1987)
10. Shaw, J.J., Bertsekas, D.P.: Optimal Scheduling of Large Hydrothermal Power Systems. *IEEE Transactions on Power Apparatus and Systems* PAS 104, 286–293 (1985)
11. Ferreira, L.A.F.M., Anderson, T., Imparato, C.F., Miller, T.E., Pang, C.K., Svoboda, A., Vojdani, A.F.: Short-Term Resource Scheduling in Multi-Area Hydrothermal Power Systems. *Electric Power & Energy Systems* 11(3), 200–212 (1989)
12. Renaud, A.: Daily Generation Management at Electricite de France: From Planning Towards Real Time. *IEEE Transaction on Automatic Control* 38(7), 1080–1093 (1993)

13. Maheswari, S., Vijayalakshmi, C.: Design and Analysis of an Optimization Model by using Scheduling Algorithm for Electric Power Cycles. In: Proceedings of the National Conference On Applied Mathematics (NCAM 2010), B.S. Abdur Rahman University Chennai, pp. 160–163 (January 2010)
14. Maheswari, S., Vijayalakshmi, C.: Optimization Model for Electricity Distribution System Control using Communication System by Lagrangian Relaxation Technique. *CiiT International Journal of Wireless Communication* 3(3), 183–187 (2011) (Print: ISSN 0974 – 9756 & Online: ISSN 0974 – 9640)
15. Wang, S.J., Shahidehpour, S.M., Kirschen, D.S., Mokhtari, S., Irisarri, G.D.: Short-Term Generation Scheduling with Transmission Constraints Using Augmented Lagrangian Relaxation. *IEEE Transactions on Power Systems* 10(3), 1294–1301 (1995)

Automatic Link Generation for the RDF Dump File: A Minimalistic Approach

Arup Sarkar¹, Ujjal Marjit², and Utpal Biswas¹

¹ Department of Comp Sc. & Engg.,

University of Kalyani, Kalyani 741235, India

² CIRM, University of Kalyani, Kalyani 741235, India

{arup,sic}@klyuniv.ac.in,

utpal01in@yahoo.com

Abstract. In this new era Semantic web technologies changed the mode of information sharing on the web. In traditional web it is familiarized to publish and share data in terms of documents. These documents are connected through hyperlinks and only meant for the human use. Semantic web enables data to be machine process able with the addition of semantic annotations by means of the data in the documents. On the semantic web these semantically annotated data remains as data silos without any knowledge and as a consequence it makes them relatively less exploitable. A contemporary technology like Linked Data is used to make them linked. Within linked Data web also called Web of Data, the data elements get connected through the links between different internal and external datasets. So, discovery of links among the entities from different datasets represents an important issue to be resolved. Our approach in this paper is to find out a simplistic way for link discovery with minimal effort. Point to be noted that without any link to any of the external data sets does not ensures a dataset is a linked data at all.

Keywords: Linked data, Rdf Dump, Web of data, Semantic Web.

1 Introduction

Linked Data [1], [2] is the medium to achieve the full usability of the proposed semantic web vision as completely working. Semantic web makes the data on web machine process-able and consumable. In general terms web is a place where data can be stored, searched, retrieved and used through some connected documents. Semantic web [3] has the potential to transfer this data documents into machine process able, understandable and semantically annotated data dumps. But the data in the dumps remains as a collection of data silos. The use of Linked Data actually makes these data silos connected. So the vision of global data space (machine process able as well as understandable) becomes possible.

1.1 Background

The presented work that is link generation is basically based on our previous work [4] where a framework is introduced to generate the linked data from the legacy database.

The job division of the whole framework is divided into three steps. First the RDF [5] dumps file generation for the data available from the legacy database. Second provenance data addition with the generated RDF dump. Third the preparation of the final Linked Data representation of RDF dumps. A more detailed description of the LGLP framework is given below in the section 2.

This work is related to the third step where actual RDF dump get ready for the linked data web. Most of the times the data Extracted from the legacy database in form of RDF dump are not connected with any of the external datasets. So formally they don't suit well to the linked data concept, since they still acts like the semantically annotated data silos. Our approach is to handle this issue. The aim is to generate a minimum number of links to some relevant external datasets with a minimal effort so that a generated RDF dump get connected with some external datasets and become ready for the web of data as a complete linked dataset.

2 Introduction to the LGLP Framework

LGLP is an effort to generate the structured linked data by extracting information from legacy data base and representing them in RDF format. The whole framework is shown in the following figure 1. Referring to the figure the framework is consists of the following components,

1. **DB:** it represents actual legacy database from which data will be extracted and converted into the machine process-able and understandable format i.e. in RDF format.
2. **RDF Dumper:** its job is to generate the actual RDF data dump. It internally uses the D2R [6], [7] map processor to generate a database-to-RDF mapping file. Next it itself processes the mapping file and generates the RDF dump. That's why it is called the RDF dumper.
3. **RDF dump:** it is the original semantically annotated data dump that we will use in our present work.
4. **Provenance Handler:** it is the component where the provenance [8] related issues will be handled. It generally generates and publishes provenance data for the dump file. We will not talk much about this component except the following points about the provenance:
 - a. States about the origin of the data.
 - b. States about the access permission of the data.
 - c. Assures about the quality of data.
 - d. Describes the data's trustworthiness.
 - e. Gives information about the used vocabularies.
5. **VOID description:** it is nothing but the provenance data represented in structured format using VOID [9], [10] vocabulary.
6. **RDF-to-HTML:** here RDF data is also represented in HTML format so that if any one tries to access any resource through traditional web browser human consumable information will return.

7. **LD Publisher:** basically component 6 is just a part of the component 7 where the actual Linked data is prepared to be published. In context of our present work, this component is the most important for us, because it is the place where the links to external datasets is measured automatically and added with original data dump to prepare them completely suitable for the Web of Data.

It is to be noted that in the original paper where the following framework is described is not introduced with the name LGLP which basically an acronym for “**Linked Data Generation from Legacy Database with provenance**”. The name is given later.

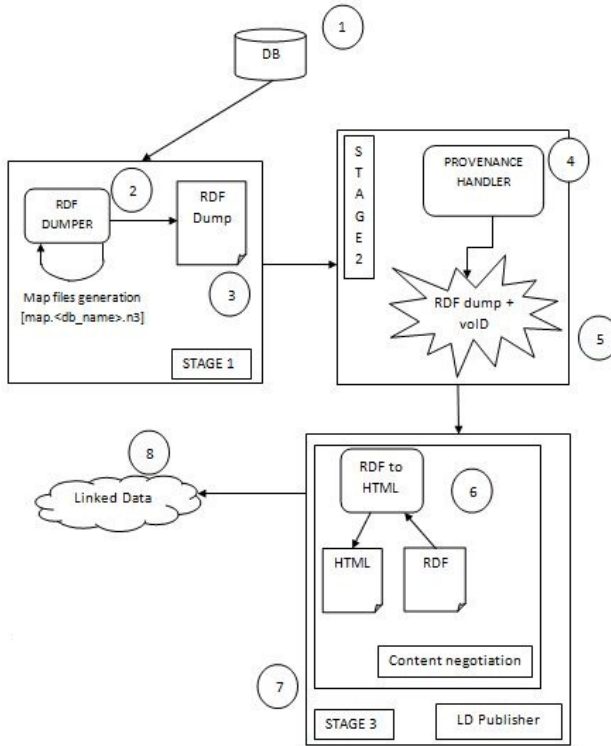


Fig. 1. Schematic Diagram of the LGLP Framework

3 Link Generation for LGLP Framework

In brief our proposed link generation approach is consists of the following five steps as shown in the following figure 2.

1. First a property file will be generated. We consider the .PROPERTIES extension for it which is common for the properties files used in a java environment. The structure of this property file is very simple. It generally holds some predefined URIs for the future use to generate the links to their corresponding datasets. It also declares a variable holding the total

- number of declared dataset URIs. This property file is generated for one time and updated if needed and used again and again.
2. Next step is to read this property file to keep the track of the predefined URIs within memory and prepare the base URIs for the external datasets.
3. Now it is time to read the actual RDF dump file and extract the concepts/ attributes/ instances/ literal values to prepare the target lists.
4. Preparation of the actual URIs, by concatenation of the base URI and the target entities.
5. Finally the link generation through the URI checking.

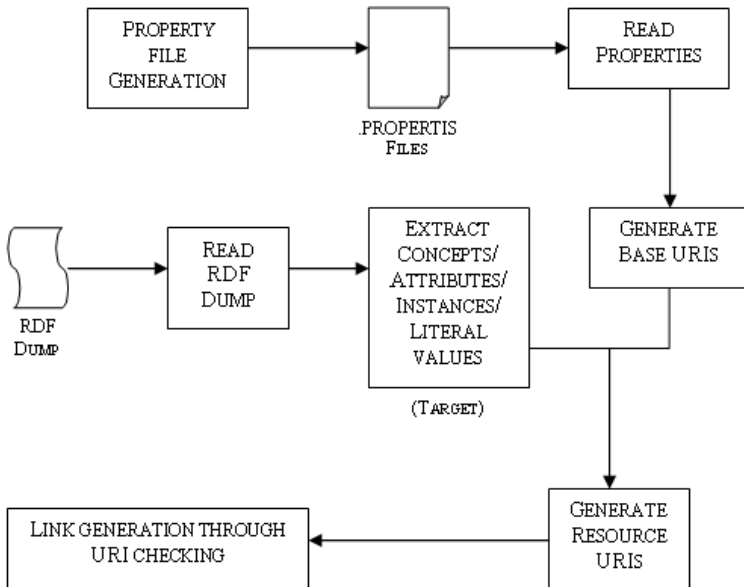


Fig. 2. Basic Block Diagram of the Internal Work flow.

Here three novel algorithms has been introduced to carry out the above mentioned operations as shown in block diagram of Fig. 2. The algorithms are named according to their jobs as “**ReadProperty**”, “**URICheck**” and the “**LinkDiscovery**”.

This first algorithm is used to read the content of the **.PROPERTIES** file, where as second one which is “**URICheck**” performs a checking on some predefined datasets with the provided resource URIs to find out if there are any related information exists about current resource. **LinkDiscovery** internally call these two algorithms to complete the link generation procedure.

ReadProperty. For the sake of simplicity of the job, **.PROPERTIES** file is used, which is a collection of key-value pairs. There are two types of key values proposed. One for holding total number of predefined dataset URI prefixes for resources and, other(s) for the actual URI prefixes for the resources on the datasets. A typical structure for the property file is shown in the following Table 1.

Table 1. Sample Property File

```

urilength = 5
uri1 = http://dbpedia.org/resource/
uri2 = http://data.linkedmdb.org/resource/country/
uri3 = http://data.linkedmdb.org/resource/film/
uri4 = http://revyu.com/things/
uri5 = http://revyu.com/people/

```

Use of this property file within the framework is very limited. Though the property file content can further be refined to hold more advance configuration related information. At present it only used as a common gateway where easily new dataset's resource URI prefixes is possible to add.

Table 2. Pseudo code for the algorithm to read a property file**Algorithm 1: ReadProperty(P)**

Input: P holds the path to the property file.

Output: U holds the list of all dataspace URIs to be searched for link discovery.

1. Read urilength from P into *len*.
 2. While $i < len$ repeat steps 3 to 6.
 3. $t \leftarrow \text{concat}(\text{uri}, i + 1)$.
 4. Read value of t from P: $u \leftarrow t$.
 5. U.add(u) ## add the value of u in to the list U ##
 6. Increment i by 1.
 7. Return U.
-

URICheck. The purpose of the algorithm is two-fold. First one is the checking of the existence of a particular resource within a predefined datasets from the properties file. On the other hand to link generation for the related resources found on the external datasets and enrich the original RDF dump with this information. The pseudo code for the algorithm is given in the Table 3.

Table 3. Pseudo code for the algorithm to check the URIs**Algorithm 2: URICheck(R, B[], T[])**

Input: R is the RDF dump, B[] is array of all the URIs of corresponding dataspace and T[] holds all the entities for which the link to be searched.

Output: R^u represents the updated RDF dump.

1. Set $i, j = 0$, $R^u \leftarrow \text{null}$.
 2. For $i = 0$ to B.length,
 - a. For $j = 0$ to T.length,
 - i. $u' \leftarrow \text{concat}(B[i], T[j])$.
 - ii. Check u' exists at B[i] or not.
 - iii. If u' exists: generate $l_n \leftarrow [<\text{rdfs:seeAlso}> \text{ link to the dataset}]$.
 - iv. $R^u \leftarrow \text{update}(R, l_n)$ for T[j].
 3. Return R^u.
-

LinkDiscovery. Previously stated two algorithms are called through the following algorithm. It takes the original RDF dump as input and returns the updated dump file enriched with new links to the external data sets. The first job of this algorithm is extraction of the concepts/instances as list of targets. Later it calls the ReadProperty and the URICheck to perform the actual link generation operations. The pseudo code for this algorithm is given below in Table 4.

Table 4. Pseudo code for LinkDiscovery algorithm

Algorithm 3: LinkDiscovery

Input: R represents the RDF dump, u[] is array of all the URIs of corresponding data spaces to be searched.

Output: R" represents the updated RDF dump with new links (specifically <rdfs:seeAlso> links).

1. Read RDF dump in R.
 2. Extract all concepts from R into C.
 3. Extract all instances from R into I.
 4. Execute ReadProperty().
 - a. Read all URIs in U.
 - b. Read U.length: urilen \leftarrow U.length.
 5. Store: B[] \leftarrow U.
 6. Store: T[] \leftarrow C.
 7. Execute URICheck() for C:
R_{temp} \leftarrow URICheck(R, B[], T[]).
 8. Change values of T[],
 - a. T[] \leftarrow null.
 - b. T[] \leftarrow I.
 9. Execute URICheck() for I:
R" \leftarrow URICheck(R_{temp}, B[], T[]).
 10. Return updated RDF dump R".
-

4 Discussion and Explanation

4.1 Scope

The LGLP framework is designed to publish the information available within any legacy database as linked data on the Web of Data. Link discovery just a part of this framework, comes with very straight forward responsibility of link generation among datasets. The present aim is to develop a mechanism with minimal facilities of link generation and limited functionalities. Concerning this aim, the concept of properties file is included, where a few hand selected well known external dataset prefixes(for resources) is declared for future use. Using this property file any dataset is possible to convert into compatible linked datasets through new link generation with minimum amount of effort. To perform this any deeper knowledge about the system active behind the scene is not necessary. The .PROPERTIES file is chosen for dataset prefixes because it is easy to read and update, which increases its reusability. It is expected that the property file will be updated with new URI prefixes and the broken one will be removed time to time.

4.2 Limitations

As already mentioned it is applicable within a limited environment with limited functionality. Number of datasets on the Web of Data is increasing day by day. Any one of them may contain related data for the entities declared into dataset under experiment. If we consider this large collection of external data spaces on the web, use of few hand picked well known dataset as target for link generation does not sound effective enough. There is a huge probability that numbers of dataspace remain unselected having really some related information and a sure destination to link up. Further, the proposed approach cannot assure that a minimum number of links to the declared datasets will be found.

4.3 New Possibilities

To make the proposed approach more dynamic and effective use of RDF crawlers is beneficial. For example Ldspider is very useful to crawl the Web of Data, searching the Web of Data for relevant information to generate the links towards them. This way dependency on some hand picked data spaces will reduce and the hit ratio corresponding to related information finding will increase.

4.4 Generated Links and Their Effects

Since the proposed mechanism is at its premature and experimental age, there is a lot of scope for development. It is reflected into the statements that we choose to generate the links. At present only `<rdfs:seeAlso>` statements are used to generate the links. In future inclusion of `<owl:sameAs>`, `<foaf:based_near>` etc. different kind of statements to be used with a little bit of modification to the above mentioned algorithms.

5 Conclusion

The proposed approach for automatic link generation is still at the budding stage. During the experimentation we have found several limitations that already been discussed in the previous section. Though, it is the most simplistic approach that may be used with minimum effort to transform any semantic data silos into a compatible linked datasets.

References

1. Berners-Lee, T.: Linked Data – Design Issues., <http://www.w3.org/DesignIssues/LinkedData.html> (accessed on May, 2011)
2. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data – The Story So Far. *Int. J. On Semantic Web and Information Syatem* 5(3), 1–22 (2009)
3. Hawke, S., Herman, I., Prud'hommeaux, E., Swick, R.: W3C Semantic Web Activity, <http://www.w3.org/2001/sw>

4. Sarkar, A., Marjit, U., Biswas, U.: Linked Data Generation for The University Data From Legacy Database. *Int. J. Of Web & Semantic Tech.* 2(3), 21–31 (2011)
5. Resource Description Framework (RDF), <http://www.w3.org/RDF/>
6. Bizer, C., Cyganiak, R.: D2R Server – Publishing Relational Databases on the Semantic Web (Poster). In: 5th Int. Semantic Web Conference, Atlanta (2006)
7. Bizer, C.: D2R Map – A DB to RDF Mapping Language. In: 12th Int. World Wide Web Conference, Budapest (2003)
8. Hartig, O.: Provenance Information in the Web of Data. In: Bizer, C., Heath, T., Berners-Lee, T., Hausenblas, M. (eds.) *Proceedings of the WWW 2009 Workshop on Linked Data on the Web*, Madrid (2009)
9. Omitola, T., Zuo, L., Gutteridge, C., Millard, I., Glasher, H., Gibbins, N., Shadbolt, N.: Tracing the Provenance of Linked Data using VoID. In: *Proceedings of the International Conference on Web intelligence, Mining and Semantics (WIMS 2011)*. ACM, New York (2011)
10. Omitola, T., Gutteridge, C.: voidp: A vocabulary for Data and Dataset Provenance, <http://www.enacting.org/provenance/voidp/> (accessed on May, 2011)

A Font Invariant Character Segmentation Technique for Printed Bangla Word Images

Ram Sarkar¹, Samir Malakar², Nibaran Das¹, Subhadip Basu¹,
Mahantapas Kundu¹, and Mita Nasipuri¹

¹ Dept. of Computer Science & Engineering,
Jadavpur University, Kolkata, India

² Dept. of Master of Computer Application,
MCKV Institute of Engineering, Howrah, India
{raamsarkar,malakarsamir,nibaran,bsubhadip,
mahantapas,mitanasipuri}@gmail.com

Abstract. A solution for segmentation of Bangla word images, printed in different fonts with varying styles and sizes, into constituent characters is reported here. Firstly, three horizontally non-intersecting zones viz., Upper, Middle and Lower Zones of a given word are identified. Then, estimation of the probable black pixels, which constitute common Matra of the word, a prominent feature in Bangla script, is done. Some of the black pixels on the Matra region are selected as potential segmentation points to segment the word vertically into their constituent characters. Each of these segmented components is then categorized into any of the six possible component types (viz. upper/middle/lower zone component/ middle and lower zone component/ broken character component/noise component). Middle and lower zone components are separated horizontally. The methodology is tested on 1600 word images of different fonts with varying styles and sizes and average success rate achieved is 96.85%.

Keywords: Character segmentation, Printed document, Bangla script, OCR.

1 Introduction

Optical Character Recognition (OCR) system contributes to the technological advancement by providing software systems to automatically convert large volumes of information, available in the form of paper documents, into electronic versions. Printed OCR systems have many applications like automation of data entry into computer from paper documents, desktop publishing, library or other office documents cataloguing etc. The tasks of OCR system can broadly be divided into four basic steps: text line extraction, word extraction, character segmentation i.e., segmentation of words into constituent characters and finally character recognition. Proper working of an OCR system for text documents highly depends on proper segmentation of words because each segment produced in this process is a candidate character prior to recognition. Obviously, the more is the accuracy of segmentation, the less will be the error during recognition process. Researchers have observed that a

large number of recognition errors in OCR system take place due to incorrect segmentation of words into constituent characters [1].

Segmentation of word images of English text into characters is usually done by identifying the valleys in the vertical pixel density histogram of word images. It is possible since consecutive characters in English text are mostly vertically separable which is not valid for the words of Bangla script. Bangla ranks 5th in the world and 2nd in India as a script and language both. It is the national language of Bangladesh. In India, Bangla is mostly used in West Bengal, Tripura and Assam. In the present work, we have developed a character segmentation technique for printed Bangla words.

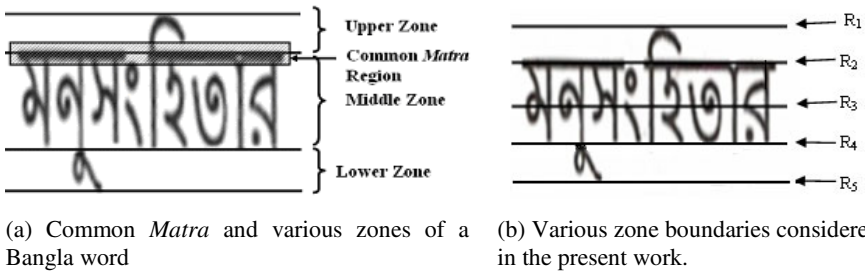


Fig. 1. (a-b). Illustration of common Matra and different zones in a Bangla word image

Bangla script consists of 50 Basic characters, 12 Modified shapes, 194 Compound characters, in addition to some special symbols including certain punctuation marks and diacritics. The vowels and some consonants take different shapes other than their original ones when they get joined with other consonants. These shapes are called modified shapes. Most characters in Bangla script have horizontal lines at the top, called the headline or Matra as shown in Fig.1 (a). Depending on character the Matra covers partly or fully the entire character width. The consecutive characters, which have Matras, are joined through a common Matra.

Any word of Bangla script can be partitioned horizontally into three adjacent zones as shown in Fig.1 (a). The portion of each word on and above the Matra is identified as the 'upper zone'. The main body of the characters in a word and the portion of the word below the main body are identified as the 'middle zone' and the 'lower zone' respectively. The characters/modifiers or part of them, which appear in the upper zone, are called ascendants, where as those in the lower zone are called descendants.

1.1 Related Review

Comprehensive surveys of strategies for character segmentation were prepared by Casey and Lecolinet [2] and by Elliman and Lancaster [3]. A brief survey on the same can also be found in the paper [4] by Arica and Yarman-Vural.

A number of works [5-10] is found in literature on segmentation of words into constituent characters for printed Bangla script. In the work by M. Chowdhury et al. [5] the Matra region was detected first and then the three basic zones of the Bangla words are identified. The authors also provided a rule-based solution to segment some

characters viz., ঞ, গ and ঞ. In [6], a complete OCR system was designed for documents written in single Bangla font. An OCR system was proposed in [7] that could read two Indian language scripts viz., Bangla and Devanagri. A complete OCR system for documents of single Bangla font was also found in [8].

Authors, in [9], presented a survey report on script segmentation for Bangla OCR and they concluded that each character from a word could be isolated in rectangular region. M. A. Hasnat et al. [10] presented a dissection based lower zone modifier segmentation method to segment the vowel modifiers present in the lower zone of a word in a wide range of document image.

In computer-composed Bangla script some characters, especially the modifiers in a word may partially overlap with one another. These kinds of nontrivial combinations of characters make the whole process of character segmentation extremely challenging. Besides, some characters / modifiers, like Chandra-Bindu (◌্ৰ) or Reph (◌্ৰ) in Bangla script, often appear in upper zone, come in between two consecutive characters in a word and isolating them from word images becomes a tough job. A variety of fonts exists for printed Bangla script. Even a text document of single font, characters may be printed in various styles (bold, italic) and / or sizes. Also proper segmentation of the ascendants and descendants from the main body of the character in the Bangla script is a tough job, which is not mentioned explicitly by most of the researchers. To overcome the above-mentioned issues, we have developed a character segmentation technique for the printed Bangla script words.

2 Present Work

A font, style and size invariant technique for segmentation of printed Bangla word images into constituent characters is presented here. After acquisition and binarization of word images, the Zone Boundary Detection module identifies three horizontally non-intersecting zones viz., Upper, Middle and Lower Zones of a given word. Then, identification of Matra region module estimates the probable rows of black pixels, which constitute the common Matra of a word. After that, some column positions on the Matra region are selected as potential segmentation points, which are subsequently used to segment the word vertically into their constituent characters or their sub-parts. Each of these segmented components are separated using Connected Component Labeling (CCL) algorithm [12] and then categorized into either of the six possible component types (viz. upper/middle/lower zone component/ middle and lower zone component/ broken character component/noise component) using a rule-based classification technique. Middle and lower zone components are separated horizontally. Detail of the present work is described in the following subsections.

2.1 Input Word Image

In the present work, we have collected isolated word images of the Bangla script from different types of documents printed in various fonts, styles and sizes of the script. Each such image is scanned using a flatbed scanner with 300 dpi resolution. The documents are then binarized by simple adaptive thresholding technique, where the threshold is chosen as the mean of the maximum and minimum gray level values in

each document image. To remove a noisy pixel and to smooth the contours of data, we have used a sequence of erosion and dilation, two basic mathematical morphological operators [12], on the input printed word images.

2.2 Zone Boundary Detection

The common *Matra* appears at the top of the main character body and acts as a boundary between upper and middle zones. For each word image, we have identified the top row of the upper zone (R_1), the top row of the middle zone (R_2), the middle row of the middle zone (R_3), the bottom row of the middle zone (R_4) and the bottom row of the lower zone (R_5) as shown in Fig.1 (b). The algorithmic representation of the method for estimation of R_1 , R_2 , R_3 , R_4 and R_5 is described below:

1. Scan the word image row-wise from top to bottom and mark the first row with one or more black pixels as R_1 .
2. Scan the word image row-wise from bottom to top and mark the first row with one or more black pixels as R_5 .
3. Scan the word image row-wise from R_1 to R_{HALF} (i.e. R_1 to $(R_1 + R_5)/2$) and compute length of the row-wise longest run of black pixels for all black pixel positions in a row.
4. Add all such longest run values in a row and select the row, appearing first from top, with maximum sum as R_2 .
5. Calculate the sum of all transition points (TPs) between foreground and background pixels for all rows from R_{HALF} to R_5 .
6. Compute η as the average number of TPs in the lower half of the image.
7. Scan the word image row-wise from R_5 to R_{HALF} and consider first row with TPs greater than η as first approximation of the bottom row of the middle zone (say R_{41}).
8. Scan the word image row-wise from R_{HALF} to R_5 and consider first row with TPs less than η as second approximation of the bottom row of the middle zone (say R_{42}).
9. The final value of R_4 is estimated as $R_4 = (R_{41} + R_{42})/2$.
10. Middle row of middle zone i.e. R_3 is calculated as, $R_3 = (R_2 + R_4)/2$.

2.3 Estimation of *Matra* Region

Matra of a printed Bangla word image may be identified as the continuous horizontal rows of black pixels in the upper half of a word image. The boundary between the sets of *Matra* pixels and non-*Matra* pixels in the region R_1 to R_{HALF} is not always distinct. The black pixels lying over the line R_2 have got strongest membership to the set of *Matra* pixels. For the other black pixels on both sides of the line R_2 , their degree of belongingness to that set diminishes as they are more and more away from the line R_2 , as per the membership function $\mu(x)$. The expression of $\mu(x)$ is given below.

$$\mu(x) = \frac{1}{1 + \left| \frac{x - c}{a} \right|^{2b}} \quad (1)$$

where 'c' denotes the center of the function, and 'a' and 'b' are parameters of the equation.

In the present work, ‘c’ is chosen as R_2 and ‘x’ as the row under consideration. The parameter ‘a’ is chosen as $(R_2-R_1)/2$ and $(R_3-R_2)/2$ for above and below the row R_2 respectively to make the function *size invariant*. In any case if the row R_1 and R_2 became collinear, the no membership value is calculated above the row R_2 . The parameter ‘b’ is chosen as 1. Estimation of *Matra* pixels, based on Eq.1, is described in detail in one of our earlier works [11].

2.4 Segmentation of Printed Word Image

In the present work, we have segmented the word images both vertically and horizontally to isolate characters and/or their sub-parts appearing in the middle zone, upper zone and lower zone of the word images.

Vertical segmentation on Matra region using fuzzy features. As *Matra* of a word image appears only in the upper half of the word image, technique of identification of vertical segmentation points is applied only on the upper half of the word images i.e. from R_1 to R_{HALF} .

One of the prominent features for identifying these points is the number of black pixels in the said region along each vertical column position on the *Matra*. The less is the number of black pixels along a vertical column position in the region (R_1 to R_2) or (R_2 to R_{HALF}) on *Matra*, the higher is its degree of belongingness to the set of segmentation points. On this basis a bell-shaped fuzzy membership function (μ_1), discussed in [11], is used. The values of parameters a, b, c are chosen as follows: $c=0$, $b=1$, $a=W_M$, where W_M is the difference between the row numbers of the two farthest *Matra* pixels in the segment, i.e., the maximum vertical width of the *Matra* region.

Another feature, the distances of a *Matra* pixel from R_2 on its both sides are considered here within the region R_1 to R_2 and R_2 to R_{HALF} . Again the more is the distance; the less is the degree of belongingness (μ_2), which is also described by a fuzzy membership function [11], of the associated *Matra* pixel to the set of segmentation points. Choices of the parameters ‘a’, ‘b’ and ‘c’ for the membership function μ_2 remain same as discussed in section 2.3, where we have applied bell-shaped fuzzy membership function (μ) to estimate *Matra* region.

For determination of final set of segmentation points, a trade-off between under/over segmentation of word images is required. For this reason, we have applied the following algorithm to identify a column (from each segmentation-point cluster) for segmentation of the word images on the *Matra* region which also minimizes the data loss due to this vertical segmentation.

1. Group the set of segmentation points, using 8-way CCL algorithm and do the following steps for each group.
2. Calculate the sum of number of data pixels, *Matra* pixels and segmentation point pixels for each column in the region R_2 to $(R_2 + (R_3 - R_2)/2)$.
3. Consider the column for vertical segmentation, which has the minimum sum.

Horizontal segmentation to separate ascendant(s) of a word. After successfully segments the word images vertically, now it is required to identify horizontal segmentation points, to isolate upper and/or lower zones character components (if connected), along R_2 and R_4 . Though in printed Bangla word images *Matra* zone is

prominent, but in some cases, the components which appear in the lower zone (i.e. the descendant), accurate estimation of R_4 becomes difficult. The reason behind this is due to presence of more number of lower zone components, italic styling of the word images or elongated vertical portion(s) of lower part(s) of the basic characters in some fonts. The descendant separation technique is described in the following subsection.

2.5 Classification of the Word Components Using a Rule Based Technique

The components generated after vertical and horizontal segmentations are classified into several classes using some threshold based rule set. This classification decision will be subsequently utilized for the reconstruction of the components to form the original word image after the recognition process is done. The classification rules use only the positional information i.e. the top-row (TR) and bottom-row (BR) of the components in the original word image. Table 1 describes the rules applied in this classification procedure.

Table 1. Different classes of word components after vertical and partial horizontal segmentation

Class #	Class Description	Classification Rule
1	Upper zone component	$(TR < (R_1 + R_2)/2)$ and $(BR \leq R_2)$
2	Middle zone component	$((R_3 - TR) > 0)$ and $(BR \geq (R_3 + R_4)/2)$ and $(BR \leq R_4)$
3	Lower zone component	$(TR > (R_3 + R_4)/2)$ and $(BR > R_4)$
4	Middle and lower zones component	$((R_3 - TR) > 0)$ and $(BR > (R_4 + R_5)/2)$
5	Broken character component	$((R_3 - TR) > 0)$ and $(BR \leq R_3)$ and (height of the component $(BR - TR) \geq (R_4 - R_2)/2)$
6	Noise component	Otherwise.

Special consideration for class #4 components. In the above-mentioned threshold-based technique, some of the class #4 components misclassified as class #2 components. To handle this problem, we have taken a special consideration for these components for possible horizontal segmentation along R_4 and subsequent re-classification of the components. The procedure is implemented in the following way.

1. If the component is a class #4 component, estimate the number of TPs i.e. changeover between foreground and background pixels and vice versa in each row starting from the middle row of the component (i.e., $(TR + BR)/2$ to R_5).
2. Select the minimum of all the TPs (say, TP_1).
3. Scan the component again in row-wise manner from its middle row to R_5 and select the first row with $TP = TP_1$ (say ROW_1). If all the rows have the same TP values from ROW_1 to R_5 then the component is considered as middle zone component.
4. Otherwise, a row (say ROW_2) is estimated where $TP > TP_1$.
5. Select the row, at the middle of ROW_1 and ROW_2 , i.e. $(ROW_1 + ROW_2)/2$, as the separation line between middle zone and lower zone.

3 Experimental Results

The character segmentation algorithm is evaluated on the word images of various types of fonts, styles and sizes printed in Bangla script. Four different font types namely, Mukti, Siyam Rupali, Vrinda and Verdana, with four different styles of printing namely, Normal, Italic, Bold and Italic-bold and five different font sizes (14 points, 16 points, 18 points, 20 points, and 22 points) are considered here. For each of the above mentioned font, style and size, we have taken 20 images of same words printed in Bangla script. Thus, a total of 1600 (i.e. $20*4*4*5$) word images are identified from different documents to include varieties in fonts, styles and sizes of the script. Some sample word images, which are properly segmented by the present technique, are shown in Table 2. Also, some word images, on which the technique fails at some points, are shown in Table 3. Failures cases are encircled in the figures, where sample #1 of Table 3 depicts the case of over segmentation and sample #2 of Table 3 shows the case of under segmentation. The average success rate of the present technique is computed to be 96.85% on the test set of 1600 Bangla word images.

Table 2. Examples of output images where the present technique works successfully

Sample#	Original word image	Output of the corresponding word image
1	সংকলন	সংকলন
2	পলাশীর	পলাশীর
3	মুমুর্ষু	মুমুর্ষু

Table 3. Examples of output images where present technique fails (errors are encircled)

Sample#	Original word image	Output of the corresponding word image
1	নিজের	নিজের
2	দৃশ্যমান	দৃশ্যমান

4 Conclusions

Computerization of Bangla script documents is still in its early stage. The printed documents may have lots of variations in terms of font, style and size of the scripts. A solution for segmentation of offline Bangla word images printed in different fonts with varying styles and sizes is reported here. Though the present technique produces reasonably good results, in future an attempt may be made to prevent the failure cases due to over/under segmentations, obtained through this experimentation, by some intelligent methodologies. Our future aim is also to apply the technique to other scripts, which are having the *Matra* on the top of the characters.

Acknowledgments. Authors are thankful to the “Center for Microprocessor Application for Training Education and Research” (CMATER), “Project on Storage Retrieval and Understanding of Video for Multimedia” (SRUVM) of Computer Science & Engineering Department, Jadavpur University, Kolkata, India for providing the infrastructure facilities during progress of the work. The work reported here, has been partially funded by PURSE (Promotion of University Research and Scientific Excellence) Programme, Dept. of Science & Technology, Govt. of India.

References

1. Nartker, T.: Information Science Research Institute 1993 Annual Report, University of Nevada, Las Vegas (1993)
2. Casey, R.G., Lecolinet, E.: A Survey of Methods and Strategies in Character Segmentation. *IEEE TPAMI* 18(7), 690–706 (1996)
3. Elliman, D.G., Lancaster, I.T.: A Review of Segmentation and Contextual Analysis Techniques for Text Recognition. *PR* 23(3/4), 337–346 (1990)
4. Arica, N., Yarman-Vural, F.T.: An Overview of Character Recognition Focused on Off-line Handwriting. *IEEE TSMC* 31 part C 2, 216–233 (2001)
5. Chowdhury, M.I.S., Dey, B., Rahaman, S.: Segmentation of Printed Bangla Characters Using Structural Properties of Bangla Script. In 5th International Conference on Electrical and Computer Engineering, pp. 639–643, Dhaka, Bangladesh, (2008)
6. Pal, U., Chaudhuri, B.B.: OCR in Bangla: an Indo- Bangladeshi language. *PR* 2, 269–273 (1994)
7. Chaudhuri, B.B., Pal, U.: An OCR System to Read Two Indian Language Scripts: Bangla and Devnagari (Hindi). In: 4th International Conference Document Analysis and Recognition, Ulm, Germany (1997)
8. Pal, U., Chaudhuri, B.B.: Computer Recognition of Printed Bangla Script. *Int. J. of Systems Sc. (IJSS)* 26, 2107–2123 (1995)
9. Billah, A., Abdullah, M., Khan, M.: A Survey on Script Segmentation for Bangla OCR. Dept. of CSE, BRAC University, Dhaka, Bangladesh (2004–2007)
10. Hasnat, M.A., Khan, M.: Rule based segmentation of lower modifiers in complex Bangla scripts. In: Conference on the Language and Technology, Lahore, Pakistan (2009)
11. Sarkar, R., Malakar, S., Das, N., Basu, S., Nasipuri, M.: A Script Independent Technique for Extraction of Characters from Handwritten Word Images. *Int. J. of Comp. Appl. (IJCA)* 1(23/17), 85–90 (2010)
12. Gonzalez, R.C., Woods, R.E.: *Digital Image Processing*, 1st edn. Prentice-Hall, Indian (1992)

Generative Process Planning Using Heuristic Artificial Intelligence Technique with CAD Modeling

Anand Satchidanandam

Mechatronics Dept, SASTRA University, Thanjavur, Tamilnadu, India
s.anand90@gmail.com

Abstract. This paper presents a novel technique to automate the solutions for the process planning in industries. A heuristic based artificial intelligence technique is used as the basic framework for the formulation of the process plan. An expert system is used for the development of the distinct nodes or states of the problem with the basis of CAD modeling. The CAD representation used for modeling may employ either Constructive Solid Geometry or Boundary representation. The comparison between the two representations is discussed along with the various other issues and steps that are pertaining to the generative process planning. This technique offers greater scope of diversification and also offers versatility in the process control leading to greater precision of the products produced in the industrial floor.

Keywords: Process planning, Artificial Intelligence, Expert system.

1 Introduction

Process planning is an elemental part of manufacturing industries where a detailed plan with all the information regarding the processes that are involved in the creation of a product is formulated. These include the type of operations performed on the material, the tools used, the order of the processes and the various parameters that are involved while performing the individual processes. Thus by obtaining the process plan, the time, efficiency and production capability (number of products that can be produced) to produce a particular product can be analyzed. Normally, this work is performed by a process engineer who has the proper knowledge of the various processes, tools and machines in the in manufacturing plant.

It can be seen that the work performed by the process engineer is manual and it largely depends on the skill-set and knowledge about the various machines and processes that are present in the manufacturing plant. Thus its dependency in the experience of the process engineer is high. Because of this the loss, in case of any error in the part of the process engineer proves to be really expensive. Moreover, the time which is required for the formulation of such a plan is also high. Thus, to increase the efficiency and also to reduce the time taken during this stage, automated process planning was introduced.

The recent developments in Artificial Intelligence (AI) and Expert Systems techniques has to led to its expansion to virtually all fields thus proving as a indispensable tool for solving of problems in various domains. Since the knowledge base of the expert system consists of past experience and also the knowledge of the

engineer who develops the expert system, [11] the system is capable of dealing with the problems efficiently thus providing the best possible practical solutions.

CAD modeling plays a vital role in the automation of the process plan. By incorporating CAD modeling in the automation scheme, it provides a direct path from the design of the product to the formulation of the process plan thus reducing the time as well as the manual work requirement.

2 Fundamentals of Planning

In order to understand the impact of using artificial intelligence in the process planning, it is best to understand the fundamentals of planning in general. What is planning? Planning is a constructive activity by which the means of reaching a particular goal is devised based on the given resources and constraints. This plan largely depends on the constraints, resources and the past experience when dealing with problems of the same kind. Thus proper feedback mechanisms have to be enforced so that the reports from the process or manufacturing line regarding the efficiency of the plan is obtained and stored for future references. One of the major ways in which the planning done by humans is simulated is by the usage of artificial intelligence and expert systems [1]. This paved the way for automated planning in various practical domains [2].

Manufacturing planning is employed in the manufacturing plant to co-ordinate the ideas of the designer with the constraints and resources in the plant to provide a proper functioning product. The manufacturing planning maybe further subdivided into operations planning, production planning and process planning [3].

In operations planning, the details regarding the various parameters and functioning of a particular process such as drilling, milling etc. are determined to ensure that the manufacturing process occurs smoothly. This type of planning is also called micro planning as it has a very narrow focus when compared to that of the other types of manufacturing planning [4].

Production planning has much broader focus when compared to the operations planning. It concentrates on the development of plan based on the availability of resources and the due-time. Thus, it takes the planning to a broader point of view without concentrating on the finer details of the production of the part. This type of planning is also called as scheduling.

In process planning, the various processes that are required to produce the final product from the raw materials. Thus, the planning contains the process route that the material should take in order for it to be transformed into the final product. The process planning includes the selection of various parameters like tool used, process sequence, feed rate and process conditions [5].

This paper mainly focuses on process planning which forms the crux of the manufacturing planning and how the application of automation improves the overall efficiency when constructing the plan.

3 Automation in Process Planning

Automation in process planning can be performed in two ways namely, generative process planning and variant process planning [6]. The two types of process planning vary in the level of automation that is used in them.

3.1 Generative Process Planning

Generative process planning is the type of planning in which the plan for new component is automatically created by the system. It creates the process plan from the information that is available in the knowledge base with the application of control logic which is done using formulas and algorithms by manipulating the geometry based input scheme used for translating the design input into usable computer code or format. Thus the various decisions such as the process selection, tools selection and operating conditions are determined automatically by using the control logic. The advantage of this system is that it is able generate a plan for a new component quickly and is able to cater to a wide range in the type of components.

3.2 Variant Process Planning

The variant process planning is based on the concept of group technology where the components are grouped into pre classified part families and the components process plan is developed from the standardized process plan that is created for that particular part family. This type of process planning system is useful when dealing with batch processes. Though the speed at which the process plan is developed is faster, the variant process planning system does not support a new component which does not belong to the pre classified part family. Thus it lacks the versatility to deal with all types of components and need the support of an engineer to constantly create and maintain the standard process plans.

Due to its advantages over the variant process planning, the generative process planning system is more widely used in automated plants. This paper, thus, concentrates on the automation in generative planning system.

4 Operating Framework

In order to provide an optimum process plan, the proper description of the work to be achieved is required. This is known as a representation problem in which the problem is represented in the proper form that enables the system to analyze and solve it [7].

The process problem in the industries can be represented follows:

1. Initial state- The raw material
2. Final State- The final product
3. Control logic
4. Knowledge Base

The first step which is the conversion of the problem into computer accessible form is done with the help of the CAD modeling and expert system. The expert system is used to analyze the CAD representation and convert the processes that have to be performed into nodes or steps. These steps only describe the action that has to be performed for the raw material to be converted into the final product. Thus, the initial state and the final state is received as input which is namely the size and shape of the of the product and the expert system determines the intermediate states or nodes. Now, the intermediate steps are not arranged in any particular order. The ordering and checking of the feasibility is done by using Artificial intelligence and the Expert

system's knowledge base and inference engine [8]. After this, the details regarding the various processes like tools used and operating environment are determined with the help of expert systems. The various steps are explained in the states diagram and the details of each step will also be explained in the following sections.

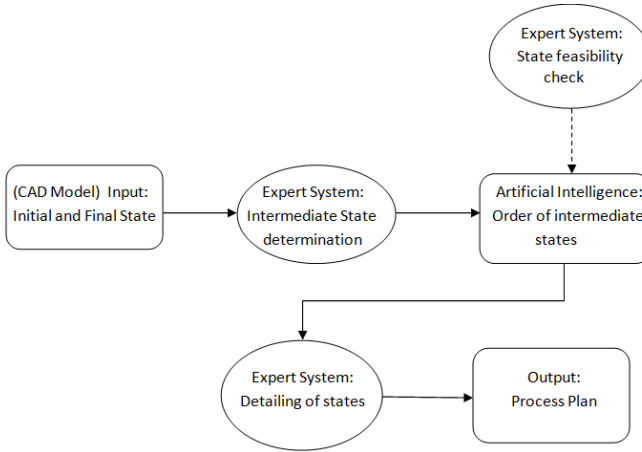


Fig. 1. Operating framework for process planning

4.1 CAD Representation

The first problem that comes into picture when dealing with process planning is the representation of the problem. There are a few characteristics that are common to all good representation systems. They should be able to fully represent all the available information in a format that is accessible by the system. The representation should be such that the structures can be easily manipulated and updated. CAD offers such a representation model which has some of the most desirable characteristics. There are two most common representations in solid modeling using CAD. They are constructive solid geometry and boundary representation.

Constructive solid geometry uses Boolean functions on basic solid structures to build the complex part. The construction of the part is stored in the form of a tree where the nodes are operations. The disadvantage here is that there is more than one way in which a solid can be created. But the advantage is that they are easy to construct and the validity of the operations can be easily checked.

Boundary representation on the other hand uses boundary surfaces as a form of representation. The advantage of this representation is that it offers a unique format for a given solid model. But the disadvantage is that it is difficult to directly incorporate this model in automated manufacturing.

Due to the ease of incorporating into the system, the Constructive solid geometry representation is considered in this model.

4.2 Expert System

The use of Expert System extends to various parts in this process planning system [9]. It is used together with the CAD representation to determine the proper order in the

Boolean combination of the basic solid structures. It is used with the artificial intelligence search technique in determining the best possible route for the creation of the final product. When finding out the best route, the various details such as the feed rate and operating conditions are also finalized using the inference mechanism of the expert system. Each of the functions will now be explained in detail.

The Constructive solid geometry representation gives the structure of the solid model as the Boolean function of the basic solid structures. But the problem with this type of representation is that the various schemes of representation in the representation tree are not checked for feasibility in terms of machining operations. For example, one cannot add or attach an external component with the main part. Only material removal is possible. Thus non-feasible schemes are removed from the representation tree by examining the entire tree using the expert system's knowledge base and inference engine [9].

Next, the expert system is used to determine the process that is required for the model to be converted from one stage to the next stage based on the solid models scheme. There are two methods in which this can be done; by forward chaining or backward chaining. In forward chaining, the expert system takes the initial stage as the base and finds the best process that might convert the model to the next stage. Since it uses the current data as a base, it is also known as data driven reasoning. In backward chaining, the goal stage is taken as the base and the process that converts the previous stage to the goal stage is determined. This is also known as goal driven. The backward chaining is used in this model of the expert system.

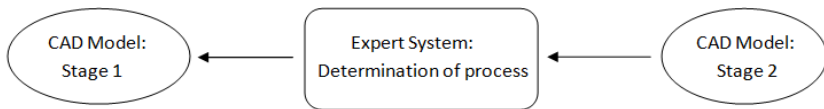


Fig. 2. Model of Backward Chaining or Goal Driven Reasoning

Now that the feasible schemes along with the processes are known, the expert system along with heuristic artificial intelligence search technique is used to determine the optimum process plan along with the details such as the feed rate and optimum working condition. The expert system is used here to help determine the optimum path based on factors such as stresses developed process, cost of process, time taken by process etc.

4.3 Heuristic Artificial Intelligence

The use of artificial intelligence in this system is mainly confined within its use in determining the optimum process scheme within the many feasible schemes that are available. Here, heuristics is used as a tool to find out if the precision and accuracy that is achieved through the process plan matches or betters that which is required in the final product by the user.

The algorithm of the artificial intelligence system should be such that it incorporates all the required functions and also completes the processing in the least possible time.

```

Start
states{}= all initial states from system
while(not empty)states{}
{cur.state=first (states{})
  Check for execution of precondition of cur.state
  Check feasibility in execution of process in cur.state
  if (feasibility not satisfied)
    {check expert system for change in process
    if(check fails)
      {remove cur.state and sub.state
      Clear fringe{}
      Update experience in expert system
      }
    }
  Update details of process from expert system
  Update heuristic function
  Compare heuristic with req. function
  if (req > heuristic)
    {check for process detail change
    if( process detail change fails)
      { remove cur.state and sub. state
      Clear fringe{}
      update experience in expert system
      }
    else
      { Change process details
      Jump to heuristic check
      }
    }
  else
    {if(sub.state{}=0)
    { fin.state[L][ ]=fringe{}
    break
    }
    else
    { states{}= insert first (sub.state)
    fringe{}=insert cur.state

    continue
    }
    }
}
Compare heuristic of fin.state[L][ ]
Print plan of fin.state with max heuristic.
End

```

Let us consider an example to explain the entire operating framework. Consider an example of a product that has to undergo drilling operation twice.

First, the problem is represented by the CAD representation and the expert system determines the feasible Boolean combination schemes. It also determines the processes that cause the transformation from one state to the other state of the scheme.

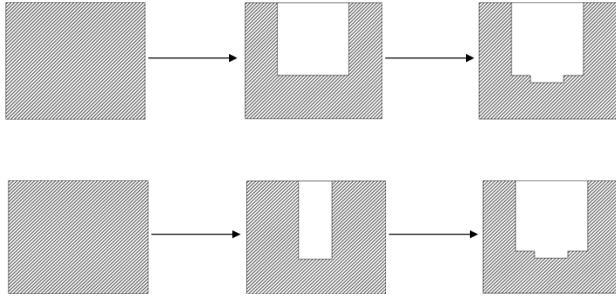


Fig. 3. Two schemes of process plan for final product

Here there are two schemes for the product. Next, the optimum process plan along with the details of the process is determined by the algorithm using expert systems.

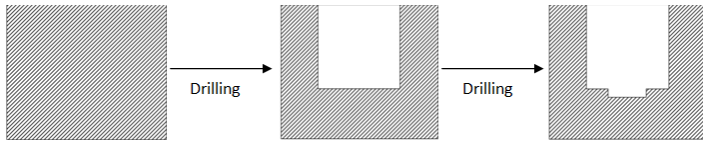


Fig. 4. Optimum process plan for final product

Here the optimum plan is that the bigger hole is drilled before the smaller one. Thus this plan has the maximum possible precision and accuracy. The plan is then sent to the manufacturing unit for production.

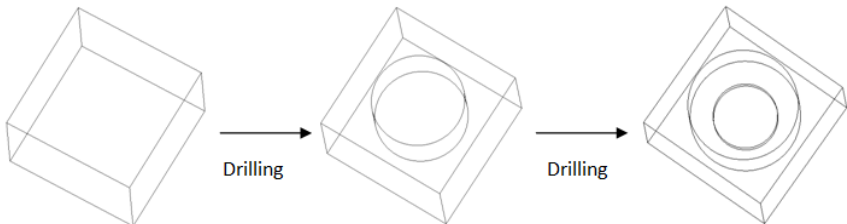


Fig. 5. 3-D process plan for final product

5 Conclusion

This paper discussed the development of automated process planning using a framework which uses expert systems, artificial intelligence and CAD modeling.

The various aspects of the framework such as the representation, the control logic and the algorithm have been reviewed. The integration of algorithmic approach of artificial intelligence with expert systems provides the means for the development of an efficient automated process plan development system which paves the way for the development of a fully automated manufacturing unit. Though there are a lot of developments in this field, there is still a lot of possibilities left to explore and thus requires more research and practical implementation. With more improvements and developments in this field, there is no doubt that the challenges of completely automating a plan will get eradicated.

References

1. Stefik, M.: Planning with Constraints. *Artificial Intelligence* 16, 111–140 (1981)
2. Feigenbaum, E.A.: *Artificial Intelligence- Themes and Case Studies of Knowledge Engineering*. *IJCAI* 5, 1014–1029 (1977)
3. Chryssolouris, G., Chan, S.: *An Integrated Approach to Process Planning and Scheduling*. *CIRP Annals* (1985)
4. Wong, T.N., Leung, C.W., Fung, R.Y.K.: Dynamic shop floor scheduling in multi-agent manufacturing system. *Exp. System App.* 31, 486–494 (2006)
5. Kals, H.J. I.: *Special contributions*, Universiteit Twente (1986)
6. Moon, C., Seo, Y.: Evolutionary algorithm for advanced process planning and scheduling in a multi-plant. *Comput. Integ. Eng.* 48, 311–325 (2005)
7. Kumar, M., Rajotia, S.: Integration of process planning and scheduling in a job shop environment. *Int. J. Adv. Manuf. Technology* 28, 109–116 (2006)
8. Shen, W.M., Wang, L.H., Hao, Q.: Agent-based distributed manufacturing process planning and scheduling: A state-of-the-art survey. *IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews* 36(4), 563–577 (2006)
9. Fuji, N., Inoue, R., Ueda, K.: Integration of process planning and scheduling using Multi-agent learning. In: *Proceeding of 41st CIRP Conference on Manuf. Systems*, pp. 297–300 (2008)
10. Li, X.Y., Zhang, C.Y., Gao, L., Li, W.D., Shao, X.Y.: An agent based approach for integrated process planning and scheduling. *Exp. System App.* 37, 1256–1264 (2010)

Simultaneous Informative Gene Extraction and Cancer Classification Using ACO-AntMiner and ACO-Random Forests

Shimantika Sharma¹, Shameek Ghosh², Narayanan Anantharaman³,
and Valadi K. Jayaraman^{2,*}

¹Department of Biotechnology and Bioinformatics, Padmashree
Dr. D.Y. Patil University, Pune, Maharashtra

²Evolutionary Computing and Image Processing Group, Centre for Development
of Advanced Computing (CDAC), Pune, Maharashtra

³Department of Chemical Engineering, National Institute of Technology,
Tiruchirapalli, Tamil Nadu, India

shimantika.sharma@gmail.com, {shameekg, jayaramanv}@cdac.in,
naraman@nitt.edu

Abstract. Microarray cancer gene expression datasets consist of high dimensional data. Gene selection helps in the removal of irrelevant genes. The reduced dimensions of the datasets help in improving the overall classification performance. We present two hybrid techniques, Ant Colony Optimization-AntMiner (ACO-AM) and ACO-RandomForests (ACO-RF) with weighted gene ranking as heuristics. The heuristic information is obtained by a weighted sum of the Information Gain, Chi-Square, Correlation based Feature Selection (CFS) and Gini Index scores for each gene. The ACO algorithm selects a small subset of relevant genes from this ranking. The fitness's of these subsets are then assessed by the *c*Ant-Miner and the Random Forest classifiers. The performances of the algorithms are tested using two cancer gene expression datasets retrieved from the Kent Ridge Bio-medical Dataset Repository. We demonstrate that genes selected by the suggested algorithms yield better classification accuracies.

Keywords: Cancer Classification, Weighted Gene Ranking, Ant Colony Optimization, *c*Ant-Miner, Random Forests.

1 Introduction

Microarray gene expression experiments allow the measurement of expression levels of thousands of genes simultaneously. This data helps in the diagnosis of various types of tumors with improved accuracy. However, one limitation of this technique is that it produces a vast amount of complex data.

* Address correspondence to the author at the Evolutionary Computing and Image Processing Group, Centre For Development of Advanced Computing, Pune University Campus, Ganesh-kind, Pune – 411007, Maharashtra, India

It is thus important to construct classifiers that can classify cancerous samples with high predictive accuracy based on their gene expression profiles.

The number of genes (features) is much greater than the number of samples in a microarray gene expression dataset. Such composition poses problems to machine learning tasks and makes the classification problem difficult to solve. This is mainly because, out of thousands of genes, most of the genes do not contribute to the classification process. To overcome this problem, one way is to select a small subset of “informative” genes from the data. This technique which is known as *Gene Selection* or *Feature Selection* not only helps in getting rid of noisy genes but also helps in reducing the computational load and in increasing the overall classification performance.

Gene selection algorithms mainly fall into two categories: *wrappers* and *filters*. Wrappers make use of a learning algorithm to estimate the quality of genes. Methods like Ant Colony Optimization [1] and Genetic Algorithm [2] in combination with a classifier like Support Vector Machine (SVM) fall into this category. On the other hand, filters evaluate the quality of genes considering the inherent characteristics of the individual genes without making use of a learning algorithm. Methods based on statistical tests and mutual information fall in this category.

This paper presents a hybrid gene selection approach which makes use of both filter as well as wrapper methods. First, genes are ranked using a weighted ranking approach (filter). Second, as a wrapper approach Ant Colony Optimization (ACO) algorithm is used [3]. ACO traverses the search space by using this ranking information coupled with pheromone mediated search to iteratively obtain more informative gene subsets. At each iteration, the selected subsets of genes, of each ant, are evaluated.

2 Methodology

2.1 Ant Colony Optimization (ACO)

Ant Colony Optimization (ACO) is an iterative search algorithm inspired by the search behavior of ant species. Real ants use an odorous chemical substance called pheromone as a communication medium for finding the food source. In search of a food source, an ant lays pheromone on its path, thus marking the path with a pheromone trail. When an isolated ant moves at random, it detects the previously laid pheromone trail and selects the path to reach the food source. In this process, this ant itself lays a certain amount of pheromone on the path, thus making the path more attractive to other ants. Thus, the probability of choosing a path by an ant increases with the number of preceding ants that chose the path.

ACO was originally used to solve the Traveling Salesman Problem (TSP) [1]. TSP is inspired by the problem faced by a salesman to find the shortest tour for which he travels through a number of cities, visiting each city exactly once before returning home. When ACO is used to solve the TSP problem, each salesman is represented by an ant. Also, in the case of gene selection, ants conduct only a partial tour in contrast to TSP [4].

ACO for gene selection mainly depends upon the following two factors:

- Heuristic information on the given gene.
- Experience (*Pheromone Information*) gathered by the ants in previous iteration.

Further details can be found in cited sources.

2.2 Heuristic Information

We obtained the heuristic information for each individual gene by calculating the weighted sum of the Information Gain, Chi-Square, CFS, and Gini index scores which were obtained using the WEKA[5] data mining software suite and Leo Breiman's Random Forests implementation[6].

Information Gain (IG) is an entropy-based measure which selects the gene that has the best capability to separate the samples into individual classes. A gene with a high information gain is said to be “informative”. Information gain is evaluated independently for each gene and the genes with the top scores are selected as the relevant genes.

Chi-square is used to measure the lack of independence between a gene and a class. Its value is closer to zero if the gene and the class are independent. We scale these scores in the range of 0-1.

Correlation-based Feature Selection (CFS) evaluates the goodness of gene subsets. CFS measures the merit of gene subsets by considering the importance of individual genes for predicting the class label along with the level of inter-correlation among them. In order to generate CFS rank scores, first the CFS attribute subset evaluator of WEKA suite was used to generate a subset of genes based on the CFS heuristic. Next, the genes selected by the evaluator were assigned a score of 0.8 and the rest were assigned a score of 0.2. These scores were used as the CFS rank scores.

Gini index is used in CART. It measures the impurity of a data partition. Like IG it selects the gene that has the best capability to separate the samples into binary partitions. So a gene with high gini index can be termed as “informative” and is subsequently ranked as per its value.

Let w_1 be the weight assigned to the Information gain (*IG*) score of gene f , w_2 be the weight assigned to the Chi-square (*CS*) score, w_3 be the weight assigned to the CFS score (*CFS*) and w_4 be the weight assigned to the Gini index (*GIN*) of the same gene. Then *Weighted Rank (WR)* of gene f is calculated as:

$$WR_f = w_1 * IG_f + w_2 * CS_f + w_3 * CFS_f + w_4 * GIN_f \quad (1)$$

2.3 Gene Subset Generation and Evaluation

An ant selects the genes based on either *exploitation* or *exploration*. For this, an algorithm parameter, q_0 ($0 \leq q_0 \leq 1$) is used to make a choice between exploitation and exploration.

If exploitation is chosen, the gene with the highest quality value corresponding to the product of the pheromone concentration on the link connecting i^{th} and j^{th} gene i.e. (f_{ij}) and the heuristic information (weighted gene ranking score) associated with the gene j i.e. $\eta(f_{ij})$ is selected. Exploration, on the other hand involves the selection of the next gene j with a probability proportional to the relative quality of the gene to the subset of genes not selected yet. This step is repeated until a partial tour of fixed gene subset size is built.

After a gene subset is constructed, each ant passes its subset to the cAnt-Miner package of Myra-2.0.1[7] tool and receives its classification accuracy. Similarly, each ant passes its subset to the randomForest 4.6-2[8] package of R and computes

classification accuracy. This accuracy is used as a fitness function for selecting the best ant of that iteration in both algorithms, executed separately.

The *cAnt-Miner* and *randomForests* classifiers' accuracies for a gene subset were evaluated using a 10-fold cross validation. The 10-fold cross validation involves breaking of data of size n into 10 sets each of size $n/10$. Out of these 10 sets, a single set is retained as the “*test data*” and the remaining 9 sets are used as the “*training data*”. The cross-validation process is repeated 10 times with each of the 10 sets used exactly once as the test data. The 10 results thus obtained are then averaged to produce a single estimation.

2.4 Pheromone Update

After each iteration the global pheromone update is performed by the best ant of that iteration. The equation for the global pheromone update is:

$$\tau_{ij} = (1-\rho) \cdot \tau_{ij} + \rho \cdot cva_{best} \quad (2)$$

where, cva_{best} is the cross validation accuracy of the best gene subset of the current iteration. The global update ensures that the desirability of those genes producing a higher accuracy is increased. This process ensures that the ants gradually learn to distinguish between the informative and the non-informative genes.

After each construction step, the local pheromone update is performed by all ants to the last edge traversed using the following equation.

$$\tau_{ij} = (1-\phi) \cdot \tau_{ij} + \tau_0 \quad (3)$$

where ϕ is in the range [0, 1] and τ_0 is the initial pheromone.

This leads to lowering the pheromone concentration on the links and hence allows the ants to choose other links thereby enabling them to produce different solutions.

2.5 *cAnt-Miner*

cAnt-Miner [9] is an extension to ACO which is used to discover classification rules. The goal of *cAnt-Miner* is to extract classification rules of the form *IF (term1) AND (term2) AND ... AND (term n) THEN (class)* from data. Each rule is made up of three terms (*attribute, operator, value*), where operator represents a relational operator and value represents a value of the domain of attribute. The rule's antecedent (*IF* part) represents the rule condition whereas the rule's consequent (*THEN* part) represents the class to be predicted by the rule.

cAnt-Miner starts with an empty rule list and adds rules one by one to the list iteratively until the number of uncovered training examples is greater than a user-specified threshold value. Ants keep adding a term to their partial rule until any term added to the rule's antecedent would make their rule cover less training samples than the user-specified value. Details can be found in the cited source.

The reduced dataset based on the gene subset formed by the previous execution of ACO is therefore provided as input to *cAnt-Miner*.

2.6 Random Forests

Random forest (RF) [6, 10] is an ensemble of randomly constructed independent decision trees. It generally exhibits substantial performance improvements over single-tree classifiers such as CART and C4.5. Randomness is introduced into the RF algorithm in two ways: one in the sample dataset for growing the tree and the other in the choice of the subset of attributes for node splitting while growing each tree. Such a RF is grown in the following manner:

- For each tree, a Bootstrap sample (with replacement) is drawn from the original training data set, i.e. a sample is taken from the training data set and is then replaced again in the data set before drawing the next sample. Likewise, ‘n’ numbers of samples are taken to form ‘In Bag’ data for a particular tree, where ‘n’ is the size of the training data set. In each of the Bootstrap training sets, about one-third of the instances are unused for making the ‘In Bag’ data on an average and these are called the out of bag (OOB) data for that particular tree.

The classification tree is induced using this ‘in bag’ data using the CART algorithm. There is no need for a separate test data in RF for checking the overall accuracy of the forest. It uses the OOB data for cross validation. After all the trees are grown, the k th tree classifies the instances that are OOB for that tree (left out by the k th tree). In this manner, each case is classified by about one third of the trees. A majority voting strategy is then employed to decide on the class affiliation of each case. The proportion of times that the voted class is not equal to the true class of case-‘n’, averaged over all the cases in the training data set is called as the OOB error estimate. The important features of random forests are that they can handle any high dimensional and multi-class data easily. Details can be found in the relevant sources as cited.

3 Results

3.1 Datasets

The output of microarray experiments are the expression levels of different genes which are available publicly. Two such datasets were obtained from the Kent Ridge Biomedical Repository [11]. The specifications of the datasets are given as per the Table 1.

Table 1. Dataset Specifications

<i>Dataset Name</i>	<i>No. of genes</i>	<i>No. of classes</i>	<i>No. of instances</i>
Colon Cancer	2000	2	62
Lymphoma	4026	2	47

3.2 Discussion of Results

The parameters and their corresponding values used in the algorithms have been listed in Table 2. These parameters were decided on after extensive evaluations and are therefore tuned to the most optimum values.

Table 2. Algorithm Parameters

<i>Algorithm Parameters</i>	<i>Values</i>
Number of ants (<i>na</i>)	20
Number of iterations (<i>itr</i>)	50
Exploitation Probability Factor(<i>q0</i>)	0.6
Pheromone Update Strength (φ)	0.25
Pheromone Decay Parameter (ρ)	0.98
Pheromone Importance Factor (β)	1
Information Gain weight (<i>w1</i>)	0.3
Chi-Square weight (<i>w2</i>)	0.1
CFS weight (<i>w3</i>)	0.4
Gini weight(<i>w4</i>)	0.2
Maximum number of ant miner iterations	1500
Number of trees in Random Forests(<i>ntree</i>)	500

Table 3. lists the sizes of gene subsets selected by the ACO-AM and ACO-RF algorithms and the 10-fold cross validations obtained for the selected gene subsets.

Table 3. 10 fold cross validation accuracies for all the datasets.

<i>Dataset</i>	<i>Original number of genes</i>	<i>Number of genes selected</i>	<i>10-fold cross validation accuracy (ACO-AM)</i>	<i>10-fold cross validation accuracy (ACO-RF)</i>
Colon	2000	5	95.47%	96.77%
Lymphoma	4026	7	96.0%	95.74%

As per our comparisons, ACO-AM and ACO-RF had outperformed the previously best performing algorithms for Colon Cancer dataset namely *SVMRFE-RG* and *Fisher-RG-SVMRFE* which had shown accuracies of 93.3 and 94.7, respectively [12-13].

Similarly the best performing algorithms for Lymphoma dataset have shown accuracies in the range of [93, 99] % earlier [14-17].

In comparison, ACO-AM and ACO-RF have shown consistent and good results.

4 Conclusions

The hybrid ACO-AM and ACO-RF methods compared well against the highest accuracies for the Colon Cancer and the Lymphoma datasets. The weighted gene ranking approach has contributed to better accuracies. The methods are simple to implement, robust and are flexible since the methods can have various possible alternatives. Parallel implementations of the proposed methods can help in obtaining the results at a much faster rate.

Acknowledgments. Dr. VKJ gratefully acknowledges the Indian National Academy of Engineering (INAE) and the Department of Science and Technology (DST), New Delhi, India for financial support.

References

1. Dorigo, M., Maniezzo, V., Colomi, A.: The Ant System: Optimization by a colony of cooperating agents. *IEEE Transactions on Systems, Man, and Cybernetics–Part B* 26(1), 1–13 (1996)
2. Goldberg, D.E.: *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Pub. Co. (1989)
3. Patil, D., Raj, R., Shingade, P., Kulkarni, B., Jayaraman, V.K.: Feature Selection and Classification Employing Hybrid Ant Colony Optimization/Random Forest Methodology. *Combinatorial Chemistry & High Throughput Screening* 12, 507–513 (2009)
4. Gupta, A., Jayaraman, V.K., Kulkarni, B.D.: Feature Selection for Cancer Classification Using Ant Colony Optimization and Support Vector Machines. In: *Analysis of Biological Data: A Soft Computing Approach*, pp. 259–280 (2007)
5. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. *SIGKDD Explorations* 11(1) (2009)
6. Breiman, L., Cutler, A. (2004), <http://www.stat.berkeley.edu/users/breiman/RandomForests/>
7. Otero, F.E.: <http://sourceforge.net/projects/myra/>
8. Liaw, A., Wiener, M.: Classification and Regression by randomForest. *R. News* 2(3), 18–22 (2002)
9. Otero, F.E.B., Freitas, A.A., Johnson, C.G.: *cAnt-Miner: An Ant Colony Classification Algorithm to Cope with Continuous Attributes*. In: Dorigo, M., Birattari, M., Blum, C., Clerc, M., Stützle, T., Winfield, A.F.T. (eds.) *ANTS 2008*. LNCS, vol. 5217, pp. 48–59. Springer, Heidelberg (2008)
10. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: *Classification and Regression Trees*. Chapman & Hall, New York (1984)
11. <http://datam.i2r.a-star.edu.sg/datasets/krbd/>
12. Mohammadi, A., Saraee, M.H., Salehi, M.: *BMC Medical Genomics*. Identification of disease-causing genes using microarray data mining and Gene Ontology, 12, doi:10.1186/1755-8794-4-12
13. Alladi, S.M., Shinde Santosh, P., Ravi, V., Murthy, U.S.: Colon cancer prediction with genetic profiles using intelligent techniques. *Bioinformatics* 3(3), 130–133 (2008)
14. Tago, C., Hanai, T.: Prognosis Prediction by Microarray Gene Expression Using Support Vector Machine. *Genome Informatics* 14(1), 324–325 (2003)
15. Li, L., Jiang, W., Li, X., Moser, K.L., Guo, Z., Du, L., Wang, Q., Topol, E.J., Wang, Q., Rao, S.: A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset. *Genomics* 85(1), 16–23 (2005)
16. Souza BF, Carvalho Ae: Gene Selection Using Genetic Algorithms. *Lecture notes in computer science* 2004, 3337(1):479-490.
17. Ziaei, L., Mehri, A.R., Salehi, M.: Application of Artificial Neural Networks in Cancer Classification and Diagnosis Prediction of a Subtype of Lymphoma Based on Gene Expression Profile. *Journal of Research in Medical Sciences* (2006)

Role Based Approach for Effective Connections in Backbone of Self Organized Wireless Networks

Neeta Shirsat and Pravin Game

Computer Engineering Department,
Pune Institute of Computer Technology,
Pune-411043, India
{neeta.shirsat,pravingame}@gmail.com

Abstract. This paper deals with the various roles that can be recognized in self organized wireless devices. Wireless network which demands for a higher level of self organization, proposes strategies of creating backbone, requires identifying various roles which each participating device is playing. In this approach the efficient way to identify the roles and devices that can perform appropriate roles suitably is identified to yield a complex global emergent behavior. Various roles are identified like agents, willingness to act as a gateway, gateways etc. A minimum connection in backbone of self organized network avoids unnecessary broadcasting and in turn energy savings can be achieved. Proposed strategy assigns roles to devices by identifying inconsistency in duplicate gateways and tries to minimize unnecessary broadcast with effective connections.

Keywords: emergent behavior, clustering, self organization, wireless devices networks, wireless communication.

1 Introduction

A Wireless, ad-hoc network is created by wireless devices like cell phones, PDAs, sensors etc provide the ability to communicate with each other directly. Various applications like disaster management, home monitoring, and office automation shows increasing demand for wireless networks. Nodes in a wireless, ad-hoc network are free to move and organize themselves in an arbitrary fashion. Physical backbone network infrastructure is not created instead nodes can communicate with each other by adapting themselves with self organization. Demands for a higher level of self organization proposes strategies of creating backbone of devices by identifying various roles which each node efficiently can perform. Devices arrive or leave in this network dynamically so it is difficult to keep all nodes connected during the changes in network topology and their local environment. Design of self organizing network communication covers various functions like design local interactions that achieve global properties , exploit implicit coordination, minimization of maintained state and design protocols that adapt to changes[1][2].

Important characteristics of self organization are emergent behavior, high level of scalability, adaptability with respect to the changes in system and robustness against

failure and damage [1][2]. Related work regarding various strategies and approaches for self organization can be found in literature which will be reviewed in next section. Various constraints on transmission ranges are considered with the work which suitably not applicable to self organization.

This paper proposes the strategy for effective connection in backbone of self organized wireless networks with role based approach. Each node is playing the efficient role for formation of backbone with local interaction. Four roles are identified: Agent, Leader, Willingness to act as a Gateways and Gateway. Each node is playing one of the roles and backbone reconfiguration can be performed with changes in environment. Network maintenance can be done with variable transmission ranges. Willingness to act as gateway role playing node avoids the problem of duplicate gateways.

The remainder of the paper is organized as follows: In section II the related work is overviewed. In Section 3 describes the algorithm for various roles assignment and the environment. In Section 4 strategy for effective connection of backbone by avoiding duplicate gateway problem and condition for change of role is proposed. Conclusions that can be drawn are covered in section 5.

2 Related Work

Energy-aware self-organization algorithms for small WSNs [3], allow to deploy a WSN solution in monitoring contexts without a base station or central nodes. Sensors are self-organized in a chain and alternate between sleep and active mode where the sleep periods are longer than the activity periods. The use of implicit coordination is exploited and shows a significant potential for reduction of power consumption. Effective creation of backbone is not considered. IDSQ ALGORITHM for Wireless sensor networks described in [4], consist of three components mainly: the sensor nodes, sensing object and the observer. Sensor nodes can be randomly placed in the human hard to reach areas in order to constitute self-organizing network, the mutual cooperation between them, each sensor node has a small processors, some data need to be addressed was sent to the node summary, then through the multi-hop routing data about monitoring on the perceived object will be sent to the gateway, and finally by the gateway to data within the entire region is transferred to the remote center to manipulate. Hence this algorithm has constraints on roles and network reconfiguration.

Various topology control approaches are designed and proposed like a multi-point relay (MPR) based approach [6], a connected dominating set (CDS) based approach [9] and a cluster based approach [5]. Flooding or broadcast storm is minimized using MPR based approach. In MPR each node maintain the one hop information and multipoint relay set(MPRS) is formed, which are responsible for forwarding the packet from a node to two-hop neighbors. CDS is used in such applications which are directly dealing with topology of wireless devices. Given an undirected graph G a subset C is a CDS of G if, for each node, u is either in C or there exists a node v such that u and v are adjacent and v is in C , and the sub graph induced by C , i.e. $G(C)$, is connected [9][10]. The nodes in the CDS are called dominators, and the others are called dominatees. But this strategy requires the knowledge of network in advance and constrained on equal transmission ranges [8]. Many cluster based approaches are proposed with self organization.

Cluster based algorithm for self organization with various roles like member, leader and gateway is proposed [5] with variable transmission ranges. Each node performs some role like member, leader or gateway and they form backbone depending on local interaction. But backbone connection can have duplicate gateway for one path which leads increase in broadcast and extra energy consumption as multiple gateways can exist on single path [5]. This scenario is shown in figure 1.

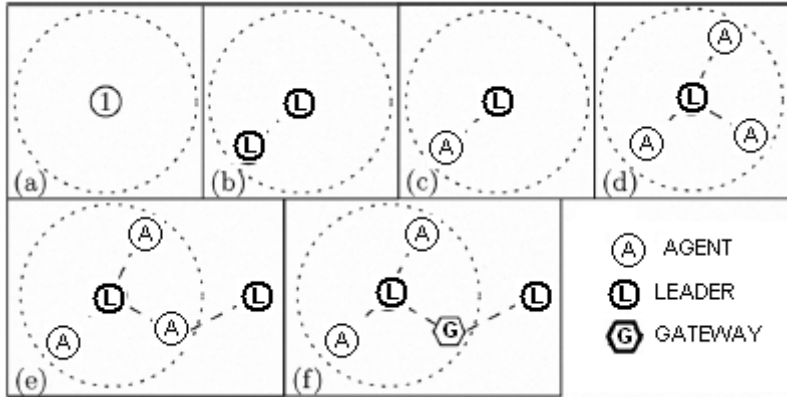


Fig. 1. Network Formation by Identifying Role

As shown in figure 1 the node which is acting as leader is connected to the all agents and leaders can communicate with the help of gateways. But the nodes which are sensing two leaders will declare themselves as gateways and will act as gateway for single path which increases the complexity and energy consumption through broadcast as shown in figure 2.

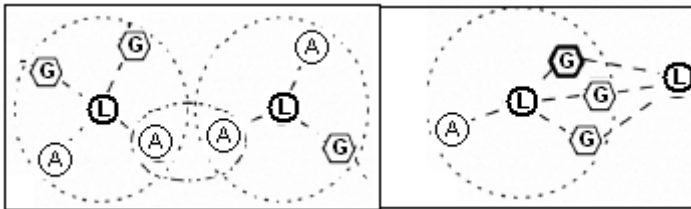


Fig. 2. Unnecessary Broadcasting

This paper proposes strategy for effective connection for backbone with a efficient gateway on single path by changing the roles of other gateways existed on same path. And tries to minimize the unnecessary broadcast and in turn energy saving is achieved.

3 Role Based Self Organization

The proposed strategy is based on self organization by identifying four different roles played by each node: Agent, Leader, Willingness to act as Gateway and Gateways.

Initially when the node wake up they do not have assigned any role. The first Role assignment is Leader for all the nodes in network. Then there is conflict for leader selection so leader election procedure is followed; so some nodes other than leaders will act as agents and some will act as gateways. This way the cluster is formed [5] figure 1.

When the new node arrives to network role is assigned first. If Leader is already exist for that group then agent role is assigned to that node else if node is detecting one or more agents without assignment of role then node role will be leader. If agent is detecting two or more leader existing in network then the role of Gateway is assigned to that node. If a node is detecting two or more leader and a gateway exist in network then the role of willingness to act as a gateway is assigned to node. At a time only one role is assigned to each node. The node assigned role as Willingness to act as a gateway will act as a normal member but it shows willingness to become a gateway to efficiently handle the failure condition and efficient connection for backbone. As willingness to act as a gateway role is assigned to such a node which can detect two groups and can work as normal agent, unnecessary broadcast is minimized and efficient backbone with a gateway on one path is maintained.

Algorithm 1. Role Based Self –Organization Algorithm

```

1: if NodeExist≠ 0
2:   if NodeLeaderNum= 0 then
3:     ROLE<=LEADER;
4:   else if ROLE =Leader then
5:     leaderElection();
6:   else if Node LeaderNum=1 then
7:     ROLE<= AGENT
8:   else
9:     ROLE<=GATEWAY
10:    if NodeGatewayNum >1 then
11:      ROLE<=WillingGateway
12:    end if
13: else
14:   ROLE<=ANY
15: end if
    
```

Algorithm 1 describes the assignment of various roles. Efficient assignment of role will definitely creates effective backbone and avoids unnecessary broadcasting. Figure 3 shows effective formation of backbone with proper role assignment.

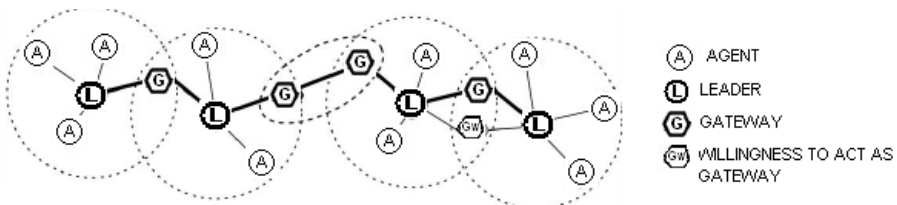


Fig. 3. Effective Formation of Backbone with proper role assignment

4 Formation of Backbone with Effective Change of Role

Initially, Modeling the environment with a connected, undirected Graph $G(V,E)$, where V is set of agents, $E \subset V \times V$ is the set of possible interconnections between pairs of agents. For each edge $(u, v) \in E$, there is a weight $w(u,v)$ that represents the cost to connect the agent u and v . We assume that we have a connected, undirected graph with $G=(V,E)$ with weight function $w : E \rightarrow R$.

Then, by using Prim’s algorithm, we find a minimum spanning tree T for $G(T \subset G)$.
Now, Consider $G = (V,E)$

$$MST \Rightarrow T = \{ V,E \} = \{ R,E' \}$$

$$\text{Where, } R = \{ L,Gw,A,W \}$$

$$\& \quad L \bullet Gw \bullet A \bullet W = \emptyset$$

(Here, R defines Roles like Leader(L), Gateways(Gw), Agents (A) and agents showing Willingness (W) to act as a gateway)

$f(R)$ is,

$$f(Gw) = Gw \text{ or } A$$

$$f(W) = Gw \text{ or } W$$

This is one implicit situation where change of role will take place.

Now, role can get changed according to energy Levels E ,

Consider,

$$\lambda = \text{minimum energy required to act as a Gateway.}$$

Let $V1' \in Gw$

Such that,

$$E(V1') < \lambda \quad \& \quad V1' \text{ is present in between } L1' \text{ and } L2'$$

$\Rightarrow V1'$ can not act as a Gateway

Decision Step:: Change Role

$$\text{Now, } \sigma(V1'' \in W) \quad \& \quad E(V1'') \geq \lambda$$

$$\& \text{ must present between } L1' \text{ and } L2'$$

Then,

$$Gw' = Gw - V1' \quad \& \quad A' = A \cup V1'$$

$\&$

$$\text{Hence } MSTB = \{ (L \cup Gw), E'' \}$$

$$\text{Where, } |E''| = |L \cup Gw| - 1$$

Now,

$$W = f(V1, L1', L2')$$

(where W is a node who is interested to be Gateway between 2 leaders, $L1'$ and $L2'$)

Role R has onto mapping such that,

$$W' = W - V1'' \quad \& \quad Gw' = Gw' \cup V1''$$

Checking Condition,

$$\text{If } V1'' \in W \quad \text{but } E(V1'') < \lambda$$

Then,

$$W' = W - V1'' \quad \& \quad A' = A \cup V1''$$

Success Factor,

$$L \bullet Gw' \bullet A' \bullet W' = \emptyset$$

Failure Factor,

$$|Gw| \neq |Gw'|$$

5 Conclusion

Role based approach for effective connections in backbone of self organized wireless networks is studied in this paper. Various roles are identified and the solution for duplicate gateway is studied. With effective connection in backbone unnecessary broadcast can be eliminated and energy saving will possible. Further study for showing the results about the proposed strategy and analyzing the energy consumption is going on.

References

1. Prehofer, C., Bettstetter, C.: Self organization in communication networks: Principles and design paradigms. *IEEE Communication Magazine* 43(7), 78–85 (2005)
2. Orfanus, D., Heimfarth, T., Janacik, P.: An Approach for Systematic Design of Emergent Self-Organization in Wireless Sensor Networks. In: *Future Computing, Service Computation, Cognitive, Adaptive, Content, Patterns, Computation World 2009*, November 15-20, pp. 92–98 (2009)
3. Kacimi, R., Dhaou, R., Beylot, A.-L.: Energy-Aware Self-Organization Algorithms for Wireless Sensor Networks. In: *Global Telecommunications Conference, IEEE GLOBECOM 2008*, November 30 -December 4, pp. 1–5. IEEE (2008)
4. Yun, B., Ji, S.-B., Li, X.: Self-Organized Algorithm Simulation for Wireless Sensor Networks. In: *2009 Second International Symposium on Information Science and Engineering (ISISE)*, December 26-28, pp. 523–526 (2009)
5. Olascuaga-Cabrera, J.G., Lopez-Mellado, E., Ramos-Corchado, F.: Self-organization of mobile devices networks. In: *Proc. IEEE Int. Conf. on Systems of Systems Engineering*, pp. 1–6 (May 2009)
6. Liang, O., Ekercioglu, Y.A.S., Mani, N.: Gateway multipoint relays- an mpr-based broadcast algorithm for ad hoc networks. In: *Proc.10th IEEE Singapore Int. Conf. Communication Systems (ICCS)*, pp. 1–6 (2006)
7. Zatout, Y., Campo, E., Llibre, J.-F.: WSN-HM: Energy-efficient Wireless Sensor Network for home monitoring. In: *2009 5th International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)*, December 7-10, pp. 367–372 (2009)
8. Correia, L.H., Macedo, D.F., dos Santos, A.L., Loureiro, A.A., Nogueira, J.M.S.: Transmission power control techniques for wireless sensor networks. *Comput. Netw.* 51(17), 4765–4779 (2007)
9. Funke, S., Kesselman, A., Meyer, M.S.U.: A simple improved distributed algorithm for minimum CDS in unit disk graphs. In: *Proc. IEEE Int. Conf. Wireless Mobile Computing, Networking, Communications (WiMob)*, vol. 2, pp. 220–223 (August 2005)
10. Nieberg, T., Hurink, J.: Wireless communication graphs. In: *Proc. 2004 Intelligent Sensors, Sensor Networks, Information Processing Conf.*, pp. 367–372 (December 2004)

Artificial Neural Network Training Using Differential Evolutionary Algorithm for Classification

Tapas Si*, Simanta Hazra, and N.D. Jana

Department of Information Technology
National Institute of Technology, Durgapur
West Bengal, India

{c2.tapas,simanta.hazra,nanda.jana}@gmail.com

Abstract. In this work, we proposed a method of artificial neural network learning using differential evolutionary(DE) algorithm. DE with global and local neighborhood based mutation(**DEGL**) algorithm is used to search the synaptic weight coefficients of neural network and to minimize the learning error in the error surface.**DEGL** is a version of DE algorithm in which both global and local neighborhood-based mutation operator is combined to create donor vector.The proposed method is applied for classification of real-world data and experimental results show the efficiency and effectiveness of the proposed method and also a comparative study has been made with classical DE algorithm.

1 Introduction

Artificial neural network(ANN) [13] is a useful tool in machine learning. ANN acts a important roll as classifier in classification of non-separable data.To apply ANN to any problem, it is necessary to train the ANN.A well-known algorithm named Back-propagation (BP) algorithm is used to train the ANN in supervise learning. BP Algorithm is a gradient descent optimize technique to search the synaptic weight coefficients of ANN and to minimize the learning error in the error surface. But BP algorithm has several drawbacks. The error function of ANN is a multi-modal function which has several local minima.The BP algorithm gets stuck into local minima easily. secondly, it has slow convergence speed.Therefore evolutionary algorithms like Genetic Algorithm(GA) [12],Particle Swarm Optimization(PSO) [6,7],Differential Evolutionary algorithm [13,4,5] are used to train the ANN as an alternative of BP algorithm. In the ANN training using GA method(GANN),weight coefficients of neural network are encoded in chromosome and selection,cross-over and mutation operators are used to minimize the error. But GANN suffers fom early convergence. GA has diversity in its population but lacks of convergence speed towards global optimia.On the other hand, PSO is applied successfully to train the ANN training[6].PSO has faster convergence speed than that of GA but it lacks of diversity in population.Recently DE

* Corresponding author.

algorithm is successfully applied for training of ANN. Advantages of DE algorithm are as follows: a possibility of finding the global minimum of a multi-modal function regardless of initial values of its parameters, quick convergence and a small number of parameters to set up at the start of the algorithm operation. In the year 2003, Fan and Lampinen [10] introduced Trigonometric DE(TDE) algorithm and applied to train the ANN as a test case for their proposed algorithm. Recently, in paper [4], DE algorithm was used to train ANN and applied to classification of parity-p problem. In Ref. [1], Adam Slowik applied adaptive DE algorithm with multiple trial vectors to ANN learning to classify parity-p problem. Liu Mingguang and Li Gaoyang combined the BP algorithm and the differential evolutionary algorithm to train the neural network in order to achieve better local search and optimizing speed in paper [3]. Yuelin Gao and Junmin Liu introduced a modified DE algorithm and trained the neural network for exclusive-OR (XOR) classification and function approximation problem in paper [5]. In this work, we trained a feed forward neural network using a DE with Local and global mutation proposed by Das et al. [2] and applied for classification of real-world data.

2 Artificial Neural Network

The n attributes in data set are used as input to NN. In this experiment, we used feed forward multi-layer perceptron (MLP, see Figure 1) that has three layers known as input, hidden and output layers respectively. Each processing node, except the input layer nodes, calculates a weighted sum of the nodes in the preceding layer to which it is connected. This weighted sum passes through the transfer function to derive its output which is fed to the nodes in the next layer. Thus, the input to node j is obtained as

$$net_j = \sum_{i=1}^M W_{ij}O_i + bias_j \tag{1}$$

and output as

$$O_j = F_a(net_j) \tag{2}$$

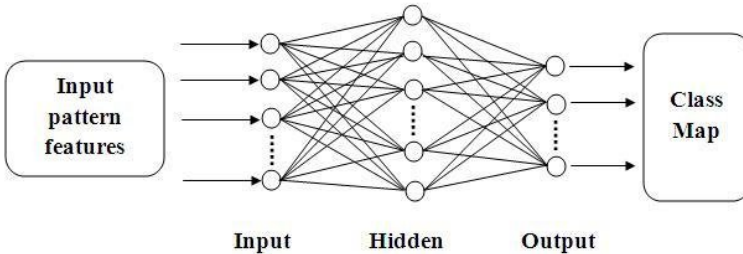


Fig. 1. Feed-forward neural network

where W_{ij} is the synaptic weight for the connection linking node i to j , value for node j , is the output of node j , and is the activation function (AF). Here the AF is considered as a sigmoid function [13] and is defined as

$$F_a(net_j) = \frac{1}{1 + e^{net_j}} \quad (3)$$

MLP uses back-propagation (BP) learning algorithm [13] for weight updating. The BP algorithm basically reduce the sum of square error called as cost function (CF), between the actual and desired output of output-layer neurons in a gradient descent manner. The CF is given as

$$CF = \sum_{i=1}^N \sum_{j=1}^M (O_{ij}^{des} - O_{ij}^{pred}) \quad (4)$$

where i is a training pattern and j is the output node. O_{ij}^{pred} denotes the predicted output of node j when the training pattern i is applied to the network, and O_{ij}^{des} is the corresponding desired output. The details of BP algorithm including derivation of equation can be obtained from [13]. The number of input nodes in the input-layer is equal to the number of attributes and the number of nodes in the output-layer is equal to the number of classes present in the data set.

3 Differential Evolutionary Algorithm

DE [8] algorithm is a floating-point population based derivative free global optimization technique. A differential operator is used to create new offspring from parent chromosomes instead of classical crossover or mutation operator in genetic algorithm(GA).In Table 1,a DE scheme, namely *DE/rand/1/bin* is described.

3.1 DEGL Algorithm

DE with local and global mutation(DEGL) is proposed by Das et al.[2].In DEGL algorithm, local mutant vector L is created using the Eq.(5)

$$L_i(t+1) = X_i(t) + \alpha_1 \cdot (X_{nbest}(t) - X_i(t)) + \beta_1 \cdot (X_p(t) - X_q(t)) \quad (5)$$

where $X_{nbest}(t)$ is the neighborhood best of i^{th} vector in iteration t and p and q are the neighborhoods of the same vector and $p, q \in [i - k, i + k]$ where $i \neq p \neq q$. k is the radius of the neighborhood of i^{th} vector in the ring topology.In this work,two neighbors are selected in the radius(k)=1 of i^{th} vector based on positional index(not geometric position). global mutant vector G is created using the Eq.(6)

$$G_i(t+1) = X_i(t) + \alpha_2 \cdot (X_{best}(t) - X_i(t)) + \beta_2 \cdot (X_r(t) - X_s(t)) \quad (6)$$

Table 1. The Main Steps of DE Algorithm

Begin
 N = population size;
 D = dimensional size;
 X = current population;
 t = the generation index;
 V = donor Vector
 U = trial Vector
 i = population index
 j = dimension index
Initialize the population of size N
while (*generation* $t \leq$ *MaxGeneration*)
for $i = 1$ **to** N
Calculate the fitness value $f(X_i(t))$
//Mutation:
for $j = 1$ **to** D
 $V_{ij}(t+1) = X_{r1j}(t) + F.(X_{r2j}(t) - X_{r3j}(t))$
// $r1, r2$ and $r3 \in [1, N]$, are integers and mutually exclusive, and $F \in (0, 2)$
// is a scale factor.
for end
//Crossover:
for $j = 1$ **to** D
if $R_j(0, 1) \leq CR$ then
 $U_{ij}(t+1) = V_{ij}(t+1)$
// $R_j(0, 1)$ is uniformly distributed random number in $(0, 1)$ and $CR \in (0, 1)$
// is crossover rate
else
 $U_{ij}(t+1) = X_{ij}(t)$
End if
for end
// Selection:
if $f(U_i(t+1)) \leq f(X_i(t))$ then
 $X_i(t+1) = U_i(t+1)$
else $X_i(t+1) = X_i(t)$
End if
for end
while end
End

where $X_{best}(t)$ is the best solution in the population in generation t . r and s are selected from $[1, NP]$ and $r \neq s \neq i$. Local mutant vector L and global mutant vector G is combined in order to create actual donor vector V using the Eq. (7)

$$V_i(t+1) = w.G_i(t+1) + (1-w).L_i(t+1) \quad (7)$$

where $w \in (0, 1)$ is a weighted factor to adjust exploration and exploitation of the search capability.

4 ANN Training Using DE Algorithm – A Review

The error function of ANN is a highly multi-modal function which has lot of local minimas. The ANN is training using well-known BP algorithm which has several drawbacks: one is that it gets stuck in local minima and another is slow convergence speed. But it has strong local search capability. The training process is to minimize the error function by adjusting weights to obtain a desired accuracy as well as to achieve faster convergence speed.

In recent few years, a lot of contributions have been given in ANN training using DE algorithm. In order to keep a reasonable balance between convergence speed and the capability of global search, Liu Mingguang et al. [3] combined the BP and DE algorithm to optimize the weights and threshold value adjustments of ANN.

Yueline Gao et al. [5] introduced a novel mutation operator in DE algorithm to obtain a good balance between global and local search and applied in BP neural network to solve exclusive-OR and function approximation problem. And as result, reduced training time and improved testing accuracy are achieved.

Adam Slowik and Michal Bialko [4] presented artificial neural network training using DE algorithm with adaptive selection of control parameters in DE and applied to classification of parity-p problem.

Adam Slowik [1] applied adaptive DE algorithm with multiple trial vectors to ANN learning to classify parity-p problem. But it takes additional training time $(m - 1) \times n_t \times G$ compare to classical DE algorithm where m is the number of trial vectors and n_t is the time taken to calculate the error function values for n records in training data set and G is the maximum generation.

Hui-Yuan Fan and Jouni Lempinen [10] introduced Trigonometric Differential Evolutionary (TDE) algorithm and they applied it to train the ANN with considering XOR problem and aerodynamic five-hole probe calibration problem.

As DEGL algorithm provides a good balance between local and global search, DEGL is used in ANN training in this work with the hope that it will provide a good performance for classification problems.

5 ANN Training Using DEGL Algorithm

In this work, ANN is trained using DEGL Algorithm (**DEGL-ANN**) to search the synaptic weight coefficients of a feed forward neural network as well as to minimize the mean-square-error in the error surface. We used a feed forward multi-layer perceptron (MLP) having n input nodes in input-layer, m output nodes in output-layer and $(2n+1)$ hidden nodes in the hidden-layer. Mean Square Error (MSE) is calculated by following Equation (8) and it is used as a fitness function for DE algorithm.

$$MSE = \frac{1}{N.M} \sum_{i=1}^N \sum_{j=1}^M (O_{ij}^{des} - O_{ij}^{pred}) \quad (8)$$

where i is a training pattern and j is the output node. O_{ij}^{pred} denotes the predicted output of node j when the training pattern i is applied to the network, O_{ij}^{des} is the corresponding desired output, N =number of training samples and M =number of outputs. For the outputs, a binary 1-of- m encoding is used in which each bit represents one of the m -possible output classes of the problem definition. Only the correct output class carries a $(1 - \epsilon)$, whereas all others carry ϵ ($= 0.1$) and winner-takes-all policy is adopted.

Total number of weight coefficients in the ANN is $D = (n * (2n + 1) + (2n + 1) * m) = (n + m)(2n + 1)$. These weight coefficients are initialized in the interval $[-1, 1]$ with uniform distribution and treated as vector elements in DE. Each and every vector in DE represents a neural network and is trained with the complete training set. After completion of maximum number of iteration or after meeting to the minimum error criteria, best neural network is used to check with the unknown test data. Finally, classification accuracy is measured by confusion matrix.

6 Data Set Description

In this work, we used five different data sets collected from UCI machine learning repository [15]. The data are normalized between $[0, 1]$ and missing values are coded as zeros. The details of the data sets are described in below:

1. Fishers iris data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other. It contains 4 attributes and one class attribute.
2. Cleveland heart disease data set has 303 records. Among the entire records, 160 are healthy, 137 are sick, and six are missing records. The data set has 13 attributes and 1 class attribute. The class attribute refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1, 2, 3, 4) from absence (value 0).
3. Breast cancer data set has total 286 instances and 9 attributes with one class attribute. 201 instances has the no-recurrence-events class whereas 85 instances has recurrence-events. This data set has missing values.
4. BUPA liver disorders data set has 345 Number of Instances and 7 attributes including one class attribute. Attribute characteristics are categorical, integer and real. This data set has no missing values.
5. Hepatitis data set has 155 number of Instances and 20 attributes including one class attribute. Attribute characteristics are categorical, integer and real. This data set has missing values.

7 Experimental Setup

K-fold cross-validation is used to obtain a reliable estimate of classifier accuracy where $K=10$ and best individual is selected in a run for testing the unknown data.

Classification accuracy was measured by Confusion Matrix [16]. The confusion matrix is a useful tool for analyzing how well a classifier can recognize tuples of different classes.

7.1 Parameters

1. Population Size(N)=30.
2. Number of Generations=100 for iris data and 200 for rest of the data sets
3. Minimum Mean Square Error (MSE)=0.005
4. $\alpha_1 = \beta_1 = 0.8$
5. $\alpha_2 = \beta_2 = 0.8$
6. CR=0.8
7. $w = 0.729$

7.2 PC Configuration

1. System:Fedora 13(i386)
2. CPU: P IV 2GHz (Core 2 Duo)
3. RAM: 3 GB
4. Software: Matlab 2010b

8 Result and Discussion

In this work, ANN is trained with both DEGL and classical DE algorithm and applied to classify the data that are described in Sect. 6. The proposed method is run with 10-fold cross-validation. The sensitivity, specificity, testing accuracy and training accuracy from best run (providing best testing accuracy) have been described in Table 2 and 3 for DEGL-ANN and DE-ANN respectively. Mean, standard deviation of training accuracy and average time of training for each data set have been described in Table 4 for DEGL-ANN and DE-ANN. Mean, standard deviation of testing accuracy for each data set have been described in Table 5 for both DEGL-ANN and DE-ANN. Convergence of mean square errors for DEGL-ANN is given in Fig. 2. The results in boldface in tables are better in the comparative analysis of two aforementioned DE algorithm. From Table 2, it is found that DEGL is able to produce a better generalization performance for neural network. From Table 4 & 5, it can be said that both classical DE and DEGL have a good efficiency in training and testing performances of neural network. The highest testing accuracy for cancer data set is 78% as per reported in UCI machine learning repository [15]. In this work, highest testing accuracy 77.19% is achieved for the same data. The highest testing accuracy is 83% as per reported in Ref. [15] for hepatitis data whereas 90.32% is achieved from this experiment. For liver data, though DE-ANN produced better testing accuracy (e.g 74.29) than that of DEGL-ANN whereas DEGL-ANN provides better mean testing accuracy.

Table 2. Best results for each data set in DEGL-ANN

Data Set	Sensitivity(%)	Specificity(%)	Testing Accuracy(%)	Training Accuracy(%)
Iris	100.00	100.00	100	97.78
Heart	89.29	93.75	91.67	85.96
Cancer	32.35	96.25	77.19	76.89
Liver	84.00	57.24	72.75	74.85
Hepatitis	100.00	40.00	90.32	95.16

Table 3. Best results for each data set in DE-ANN

Data Set	Sensitivity(%)	Specificity(%)	Testing Accuracy(%)	Training Accuracy(%)
Iris	100.00	100.00	100	100
Heart	78.94	87.50	83.57	83.92
Cancer	32.00	91.67	74.12	74.90
Liver	85.00	60.00	74.29	66.45
Hepatitis	89.66	100.00	90.21	83.06

Table 4. Mean,standard deviation of training accuracy and average time of training for each data set

Data Set	DEGL-ANN			DE-ANN		
	Mean	Std. Dev.	Avg. Time(hrs.)	Mean	Std. Dev.	Avg. Time(hrs.)
Iris	98.85	0.5040	0.005	100.00	0.0	0.007
Heart	86.46	0.3698	0.27	83.53	0.4572	0.268
Cancer	77.3123	0.2251	0.26	74.83	0.6511	0.26
Liver	75.59	0.72	0.32	70.26	1.4323	0.31
Hepatitis	92.66	2.6462	0.13	83.79	3.0756	0.128

Table 5. Mean,standard deviation of testing accuracy for each data set in both DEGL-ANN and DE-ANN

Data Set	DEGL-ANN		DE-ANN	
	Mean	Std. Dev.	Mean	Std. Dev.
Iris	98.71	0.9367	100.00	0.0
Heart	86.44	3.1712	82.79	0.5945
Cancer	73.97	2.2862	72.54	1.3943
Liver	70.67	2.1930	68.50	2.4320
Hepatitis	82.90	6.8091	81.94	5.3114

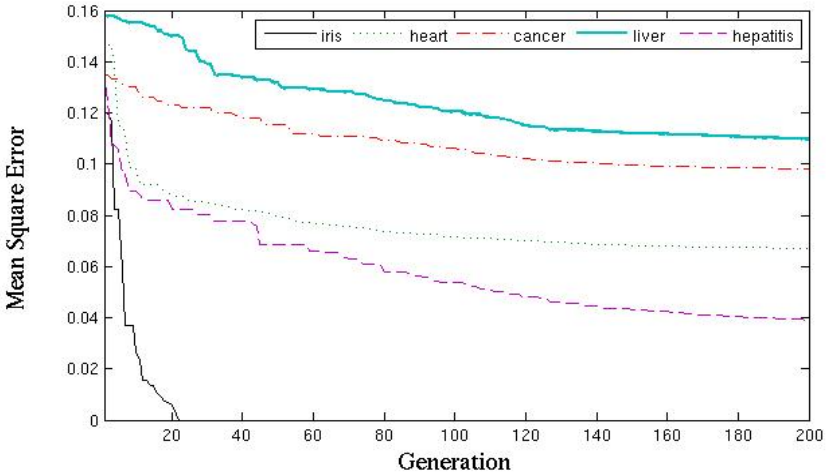


Fig. 2. Convergence graph for DEGL-ANN

9 Conclusion and Future Works

In this work, artificial neural network is trained using DE with local and global mutation and classical DE algorithm and application has been done for classification of real-world data set. From the experimental results, it has been shown that DEGL algorithm is efficient and effective in neural network learning with producing a good generalization performance than classical DE algorithm. Performances of a classifier are varied due to complexity of data set, preprocessing of data (i.e removing or replacing the missing values in data), transformation of categorical values to numerical values. Proper preprocessing scheme can be adopted to enhance the performances of the proposed method. In this work, fixed parameter values are used in all iteration of the DEGL algorithm. Different control mechanism for parameters setting in DEGL algorithm can be adopted while training the neural network. From this study, it may be concluded that DEGL algorithm can be used in ANN learning for classification problems.

References

1. Slowik, A.: Application of an Adaptive Differential Evolution With Multiple Trial Vectors to Artificial Neural Network Training. *IEEE Transactions On Industrial Electronics* 58(8), 3160–3167 (2011)
2. Das, S., Abraham, A., Chakrabarti, U.K., Konar, A.: Differential Evolution Using a Neighborhood-Based Mutation Operator. *IEEE Transactions On Evolutionary Computation* 13(3), 526–553 (2009)
3. Mingguang, L., Gaoyang, L.: Artificial Neural Network Co-optimization Algorithm based on Differential Evolution. In: *Second International Symposium on Computational Intelligence and Design*, pp. 256–559 (2009)

4. Slowik, A., Bialko, M.: Training of Artificial Neural Networks Using Differential Evolution Algorithm. In: 2008 Conference on Human System Interactions, pp. 60–65 (2008)
5. Gao, Y., Liu, J.: A Modified Differential Evolution Algorithm and Its Application in the Training of BP Neural Network. In: Proceedings of the 2008 IEEE/ASME International Conference on Advanced Intelligent Mechatronics, Xi'an, China, pp. 1373–1377 (2008)
6. Lee, Y.S., Shamsuddin, S.M., Hamed, H.N.: Bounded PSO Vmax Function in Neural Network Learning. In: Eighth International Conference on Intelligent Systems Design and Applications, pp. 474–479. IEEE (2008)
7. Junyou, B.: Stock Price forecasting using PSO-trained neural networks. IEEE Congress on Evolutionary Computation, 2879–2885 (2007)
8. Price, K., Storn, R., Lampinen, J.: Differential Evolution – A Practical Approach to Global Optimization. Springer, Heidelberg (2005)
9. Yan, H., Zheng, J., Jiang, Y., Peng, C., Li, Q.: Development of a Decision Support System for heart Disease Diagnosis using Multilayer Perceptron. In: Proceedings of the 2003 International Symposium on Circuits and Systems, pp. 709–712 (2003)
10. Fan, H.Y., Lampinen, J.: A Trigonometric Mutation Operation to Differential Evolution. International Journal of Global Optimization 27, 105–129 (2003)
11. Shi, Y., Eberhart, R.C.: A modified particle swarm optimizer. In: Proc. IEEE World Congr. Comput. Intell., pp. 69–73 (1998)
12. Tsi, D.-Y.: Classification of Heart Diseases in Ultrasonic Images using Neural Networks Trained by Genetic Algorithm. In: Proceedings on International Conference on Signal Processing, pp. 1213–1216 (1998)
13. Haykin, S.: Neural Networks - A Comprehensive Foundation, 2nd edn. PHI (1994)
14. Rajasekaran, S., Pai, G.A.V.: Neural Networks, Fuzzy Logic, and Genetic Algorithms Synthesis and Applications. PHI (2008)
15. <http://cml.ics.uci.edu>
16. Han, J., Kamber, M.: Data Mining - Concepts and Technique. Elsevier (2006)

A New Unsharp Masking Algorithm for Mammography Using Non-linear Enhancement Function

Siddharth, Rohit Gupta, and Vikrant Bhateja

Department of Electronics and Communication Engineering
Shri Ramswaroop Memorial Group of Professional Colleges
Lucknow-227105 (U.P.), India
{link.siddharth,bhateja.vikrant}@gmail.com,
grohiteng24@yahoo.in

Abstract. Mammography is especially valuable as an early detection tool because it can identify breast cancer at a stage when treatment may be more effective. This paper introduces a new Unsharp Masking (UM) algorithm using a non-linear enhancement function. The proposed algorithm combines the conventional UM with the non-linear enhancement function. The conventional UM algorithm is extremely sensitive to noise because of the presence of the linear high pass filter. The improved high pass filter used in the proposed work provides high frequency components of the image which are insensitive to noise which reduces the noise sensitivity of the UM algorithm. The input image is simultaneously processed using the improved high pass filter and the non-linear enhancement function; both the images are then combined to get the final enhanced image. Simulation results show that the proposed algorithm not only enhances the edges of the masses, but at the same time suppresses the background noise as well.

Keywords: Unsharp Masking (UM), Non-Linear Enhancement function, Digital Mammography, Region Segmentation.

1 Introduction

Cancer is a group of diseases characterized by uncontrolled growth and spread of abnormal cells. If the spread is not controlled, it can result in death. Worldwide, one in eight deaths is due to cancer [1]. According to estimates from the International Agency for Research on Cancer (IARC), there were 12.7 million new cancer cases in 2008 worldwide, of which 5.6 million occurred in economically developed countries and 7.1 million in economically developing countries[2]. Breast Cancer is the most common cancer among women (after skin cancer). According to the report of American Cancer Society, breast cancer accounts for nearly 1 in 4 cancers diagnosed in US women [2]. This cancer may be invasive if it has spread from the milk duct or lobule to healthy breast tissues. Development of a tumor is an outcome of various internal processes affecting the breast tissues in different ways. A detected tumor always needs further investigation unless it is classified one among the well-defined types. A round, oval, or lobulated mass with sharply defined borders has a high likelihood of being benign. On the other hand those with distorted borders need

further enhancement to be classified as a tumor. Masses with irregular boundaries are generally malignant [3]. Mammography is capable of detecting and locating underlying tumors which are non-palpable in nature that may be cancerous at a later stage. Mammographic images are limited by poor radiologic resolution, especially in case of patients with denser breasts, prior surgery, previous radiation or breast implants. For accurate computer aided detection of this breast tumor, edge enhancement techniques are applied to digital mammograms. Thus, computer assisted diagnostic techniques serve to be an important tool for improving breast cancer detection. These techniques are useful in early detection of breast cancer, thus providing a remarkable reduction in disease mortality [4]. Conventional enhancement techniques applicable in the spatial domain are not very effective in case of mammographic images. Histogram based enhancement techniques well preserves the edges of the masses but there are losses of details outside the denser parts of images. Unsharp Masking (UM) is a very common technique used to enhance the edges of the breast tumor. Conventional linear UM method [5] is very simple but because of usage of linear high pass filter in the linear UM method makes the system extremely sensitive to noise. In addition the usage of a global enhancement factor leads to over-enhancement thereby introducing some undesired artifacts in the finally processed image. Many variants of the conventional UM methods are proposed in the literature to overcome these limitations. Strobel *et al.* used quadratic operator in place of linear high pass filter [6] to improve the performance of UM method. However, usage of this operator introduces some visible noise depending on the enhancement factor. The properties of quadratic filters were used to generate high order polynomial operators [7] suitable for contrast enhancement using UM method. But to reduce the noise effects, the output of the high pass filter is multiplied by a control signal obtained from the output of an edge sensor. Mira *et al.* proposed a normalized non-linear approach [8] which replaced the high pass component of the conventional UM method by a fraction obtained from the quadratic filter. However, this approach amplified the noise in high contrast areas. The cubic UM approach [9] was suggested by Ramponi *et al.* which uses a quadratic function of local gradient to suppress noise. This filter works well in some areas but may introduce some visible noise depending on the choice of enhancement factor. Adaptive UM method proposed by Polosel *et al.* [10] cannot be applied for the edge enhancement of mammographic images because of the high complexity. Yang *et al.* proposed a UM method based on region segmentation [11] to overcome the drawback of adaptive UM method [10]. However, the authors have used conventional high pass filter, hence it is not able to enhance the details and the lesion edges in region of interest (ROI) effectively. Wu *et al.* proposed an improved unsharp mask method [12], but some overshoot artifacts can still be seen and the edges are also not clearly visible. This paper introduces an modified UM algorithm based on non-linear enhancement function combined with the region segmentation for digital mammograms. The present work proposes a new 5x5 template to improve the performance of the high pass filter, thus providing high frequency components of the targeted ROI which are insensitive to noise. Combination of this high pass filter with the UM based on region segmentation, not only enhances the contrast of the lesion details and edges but also suppresses the background noise. The rest of the paper is organized as follows: Section 2 discusses the improved UM based on region segmentation method in the first half. The latter

half of this section describes the non-linear enhancement function. Results and discussions are covered in Section 3, while the concluding remarks are given under Section 4.

2 Proposed UM Algorithm

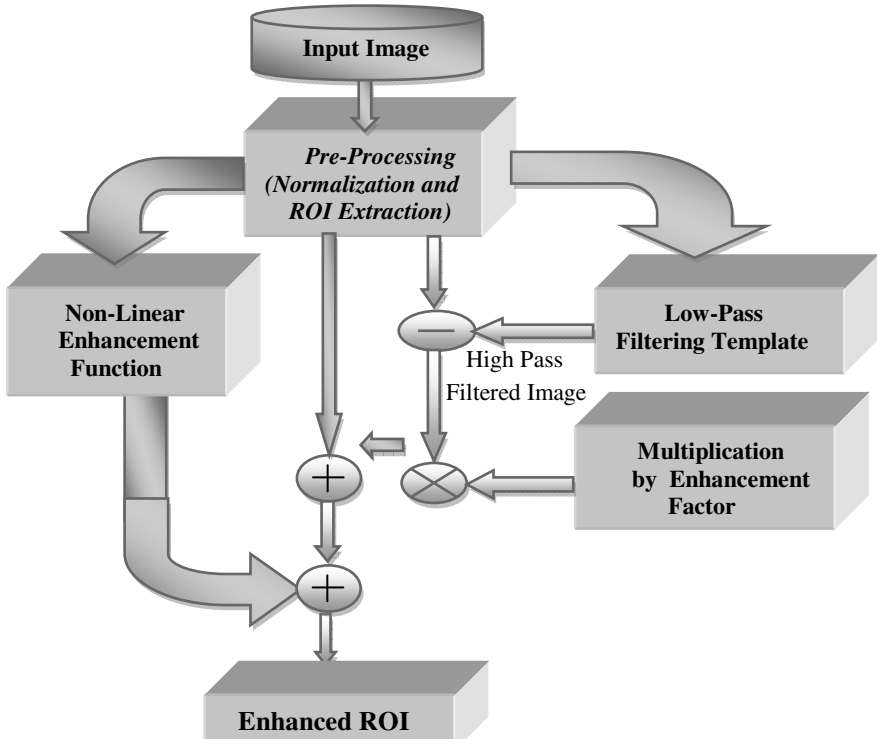


Fig. 1. Block diagram of the proposed algorithm

2.1 Improved UM Based on Region Segmentation

UM based on region segmentation [13] works on the principle of dividing the entire image into three different segments according to their characteristics. These three segments are low-detail, medium-detail and high-detail regions correspond to low, medium and high frequency regions respectively. To assign a pixel to any one of these three segments, a local variance v_i is computed over a 3x3 pixel block using the formula:

$$v_i(m, n) = \frac{1}{9} \left(\sum_{i=m-1}^{m+1} \sum_{j=n-1}^{n+1} (h(i, j) - \bar{h}(m, n))^2 \right) \tag{1}$$

where: $\bar{h}(m, n)$ is the average luminance level of the 3x3 pixel block. The conventional linear equation for UM can therefore be modified and stated as :

$$f(x, y) = c(x, y) + \beta(x, y)g(x, y) \tag{2}$$

where: $f(x, y)$ is the final enhanced image, $c(x, y)$ is the original image and $g(x, y)$ is the output of the linear high pass filter. The new convolution template proposed in this work effectively improves the high pass, thereby modifying the sharpening effect of the UM algorithm. The new high pass filter produces the high pass filtered image by subtracting the low pass filtered image from the original image. The background prediction process is achieved using the low-pass filter and then this image is subtracted from the original input image producing a high frequency image in which the background information is suppressed. The expression for the improved high pass filter can be given as:

$$g'(x, y) = c(x, y) - l(x, y) \tag{3}$$

$l(x, y)$ is the image containing the background information obtained from a low pass filter. The low pass filtering and the background prediction is performed by convolution of the proposed template $H(m, n)$ with the original image $c(x, y)$ is defined as:

$$l(x, y) = c(x, y) * H(m, n) \tag{4}$$

$H(m, n)$ is the 5x5 convolution template expressed as :

$$H(m, n) = \frac{1}{60a} \begin{bmatrix} a & a & a & a & a \\ a & 4a & 3a & 4a & a \\ a & 3a & 0 & 3a & a \\ a & 4a & 3a & 4a & a \\ a & a & a & a & a \end{bmatrix} \tag{5}$$

where: a can take whole numbers between 4 and 8. Now, substituting the output of the high pass filter $g'(x, y)$, in (2), the modified expression for the UM method based on region segmentation can be stated as:

$$f(x, y) = c(x, y) + \beta(x, y)g'(x, y) \tag{6}$$

The proposed method in (6) is applied to the ROI with low enhancement factor, in order to effectively extrude the edges of mammographic masses. The block diagram representation of the proposed UM algorithm is shown in fig. 1.

2.2 Non-linear Enhancement Function

The major difficulty in the contrast improvement of mammograms is that the noise present in the background is also enhanced during the enhancement of fine details of the ROI, this makes the tumor undistinguishable from the background. Non linear enhancement approach with multistage adaptive gain [13] is used to overcome the above problem. Linear combination of logistic function along with the gain factor is used for the preparation of the non-linear mask for the enhancement of ROI. This function modifies the gray levels by the suppression of pixel values of smaller amplitude and enhancement of only those pixels larger than a certain threshold. The non-linear enhancement function [14] used to perform the above operation is given by:

$$y(x) = a[\text{logistic}\{k(x - b)\} - \text{logistic}\{-k(x + b)\}] \tag{7}$$

where x denotes the gray level value of original ROI at co-ordinates (i,j) , k is a parameter for control of enhancement and b is the threshold value to be chosen. A logistic function is a real valued, differentiable and monotonically increasing function given by :

$$\text{logistic}(x) = \frac{1}{1 + e^{-x}} \tag{8}$$

where a is given by :

$$a = \frac{1}{\text{logistic}\{k(1 - b)\} - \text{logistic}\{-k(1 + b)\}} \tag{9}$$

where $0 < b < 1, b \in \mathbb{R}$ while $k \in \mathbb{N}$

The graphical variation of the non linear enhancement function is shown in fig 2.

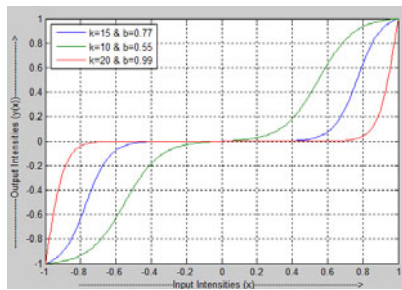


Fig. 2. Graphical variation of Non-Linear Enhancement function $y(x)$

3 Results and Discussion

3.1 Simulation Results

The images used in this work for simulations are taken from the Mammographic Image Analysis Society (MIAS) database [15] which is publically available and one of the most easily accessed databases consisting of 322 digital mammograms. The pre-processing operation involves ROI extraction, where a particular section of the mammographic image is cropped, from the probable area of lesion location and then normalized. The test images include ROI of size 256 x 256 pixels extracted from mammograms containing masses which may be benign or malignant. The proposed UM method is then applied to the pre-processed ROI. The enhancement results on the different ROI are shown in Fig. 3. In the proposed UM method, $a=4$ is used for the filter template (5). The values of k and b used in the enhancement function are 15 and 0.77 respectively. The proposed method is then applied with low enhancement factors.

3.2 Evaluation Method Used

Combined Enhancement Measure (CEM) [16] is used as an evaluation parameter for the ROI processed by different algorithms. It combines DSM , $TBCe$ and $TBCs$ for a

particular algorithm by representing each value within a three dimensional Euclidean space. The algorithm giving the smallest value of *CEM* is selected as the best enhancement algorithm for ROI. These parameters can be calculated as under:

$$DSM = (\mu_T^E - \mu_B^E) - (\mu_T^O - \mu_B^O) \tag{10}$$

$$TBC_s = \left\{ \frac{(\mu_T^E / \mu_B^E) - (\mu_T^O / \mu_B^O)}{\sigma_T^E / \sigma_T^O} \right\} \tag{11}$$

$$TBC_e = \left\{ \frac{(\mu_T^E / \mu_B^E) - (\mu_T^O / \mu_B^O)}{\varepsilon_T^E / \varepsilon_T^O} \right\} \tag{12}$$

$$CEM = \sqrt{(1 - DSM)^2 + (1 - TBC_s)^2 + (1 - TBC_e)^2} \tag{13}$$

where: $\mu_B^O, \mu_T^O, \sigma_T^O, \varepsilon_T^O$ are the mean, standard deviation and entropy of the gray scales comprising the background and the target area, of the original image before enhancement whereas, $\mu_B^E, \mu_T^E, \sigma_T^E, \varepsilon_T^E$ are the mean , standard deviation of the gray scales after enhancement. *CEM* values of the digital mammograms processed by various algorithms are enlisted under table 2.

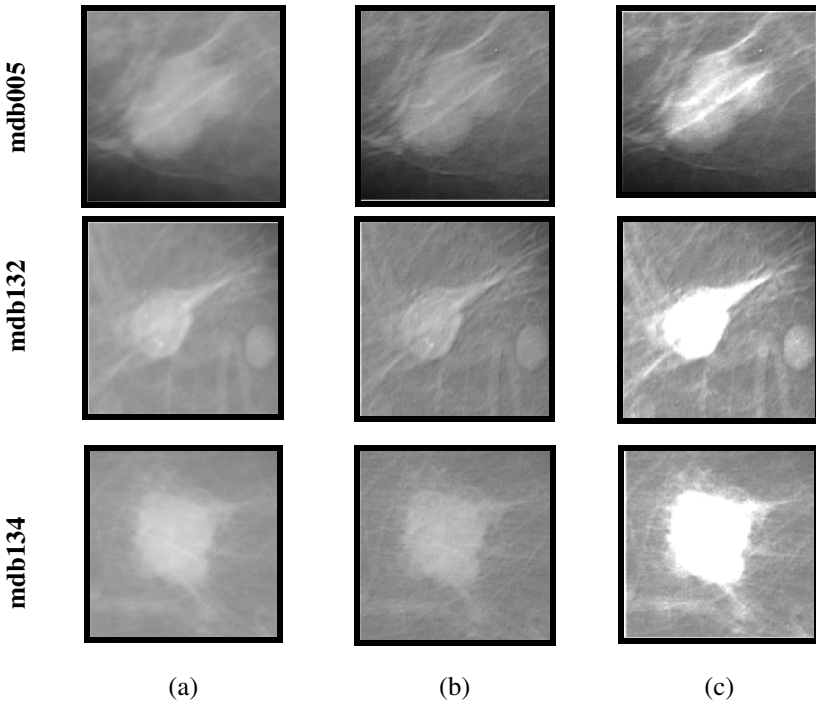


Fig. 3. Enhanced ROI images obtained with proposed methodology (a) Pre-Processed ROI (b) Enhanced ROI using low-pass filtering template (c) Enhanced ROI using proposed method

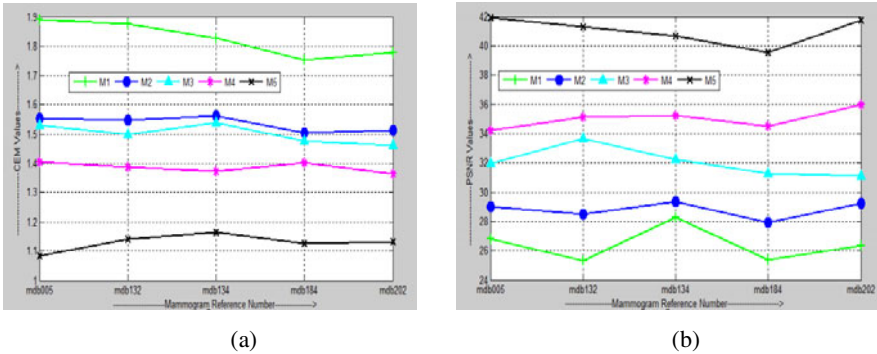


Fig. 4. Graph showing the comparison of values obtained by applying various UM algorithms. (a) *CEM* (b) *PSNR*.

3.4 Comparison of Results

It can be observed from the results obtained in table II that the Linear UM method (M1) [5] yields a high value of *CEM*. UM based on region segmentation(M2 &M3) [13] yields lower values of *CEM* in comparison to linear UM (M1) [5] but the value of *PSNR* is not appreciable/appropriate. The proposed UM method (as shown in fig. 3 (c)) produces the ROI with enhanced edges along with due suppression of background noise, thereby yielding the lowest value of *CEM* in comparison to other algorithms and the *PSNR* improves showing the increase in the overall quality of the mammogram. It can be observed from the graph in figure 4 (a) and (b) that the *CEM* values obtained from the proposed algorithm are lowest in comparison to the other UM algorithms whereas the *PSNR* values are the highest, thus proving its better performance. The results are even better as compared to the recent proposed UM method by Wu *et al.*

Table 2. Values of Evaluation Parameters of Digital Mammograms Processed by Various Algorithms

ROI	M1		M2		M3		M4		M5	
	PSNR	CEM	PSNR	CEM	PSNR	CEM	PSNR	CEM	PSNR	CEM
mdb005	26.81	1.89	28.99	1.55	31.96	1.52	34.21	1.40	41.90	1.08
mdb132	25.31	1.87	28.48	1.54	33.62	1.49	35.10	1.38	41.26	1.14
mdb134	28.28	1.82	29.35	1.56	32.24	1.53	35.21	1.37	40.69	1.16
mdb184	25.38	1.75	27.94	1.50	31.24	1.47	34.49	1.40	39.56	1.12
mdb202	26.34	1.77	29.21	1.51	31.12	1.46	35.94	1.36	41.76	1.13

M1 :Linear UM [5], M2 :RS with high enhancement factor [11], M3 :RS with low enhancement factor [11], M4 :UM algorithm proposed by Wu *et al.* [12], M5 :Proposed Method

4 Conclusion

This paper introduces a new non-linear UM algorithm for the enhancement of mammographic masses. The Linear or conventional UM algorithm suffer from drawback of being sensitive to noise because of the presence of the linear high pass

filter as well as the enhanced images contains overshoots which is an undesirable feature. The proposed algorithm uses an improved high pass filter template to enhance the performance of the high pass filter, and then combines the results obtained from non-linear enhancement function. Simulation results demonstrate that the proposed algorithm enhances the edges of the noise effectively along with the suppression of background noise.

References

1. American Cancer Society. *Global Cancer Facts & Figures*, 2nd edn. (2011)
2. American Cancer Society, *Breast Cancer Facts & Figures 2009-2010* (2009)
3. Kopans, D.B.: *Breast Imaging*, 3rd edn. Williams & Wilkins, Baltimore (2007)
4. Tang, J., Rangayyan, R.M., Xu, J., Naqa, I.E., Yang, Y.: Computer-Aided Detection and Diagnosis of Breast Cancer With Mammography: Recent Advances. *IEEE Transactions on Information Technology in Biomedicine* 13(2), 236–251 (2009)
5. Rogowska, J., Preston, K., Sashin, D.: Evaluation of digital unsharp masking and local contrast stretching as applied to chest radiology. *IEEE Transactions on Biomedical Engineering* 35(10), 817–827 (1988)
6. Strobel, N.: Quadratic Filters for Image Contrast enhancement. Dept. of Electrical Engineering, University of California, Santa Barbara (June 1994)
7. Ramponi, G., Stroble, N., Mitra, S.K., Yu, T.: Nonlinear Unsharp Masking methods for image contrast enhancement. *Electron Image* 5, 353–366 (1996)
8. Yu, T.H., Mitra, S.K.: Unsharp masking with nonlinear filters. In: *Proc. of Seventh European Signal Processing Conf., EUSIPCO 1994, Scotland* (September 1994)
9. Ramponi, G.: A cubic unsharp masking technique for contrast enhancement. *Signal Process.* 67, 211–222 (1998)
10. Polosel, A., Ramponi, G., John Mathews, V.: Image Enhancement Via Adaptive Unsharp Masking. *IEEE Transactions on Image Processing* 9, 505–510 (2000)
11. Yang, Y.B., Shang, H.B., Jia, G.C., Huang, L.Q.: Adaptive unsharp masking method based on region segmentation. *Optics and Precision Engineering* 11, 188–191 (2003) (in Chinese)
12. Wu, Z., Yuan, J., Lv, B., Zheng, X.: Digital mammography image enhancement using improved unsharp masking approach. In: *Proc. of 3rd International Conference on Image and Signal Processing*, pp. 668–671 (December 2010)
13. Laine, A.F., Schuler, S., Fan, J., Huda, W.: Mammographic feature enhancement by multiscale analysis. *IEEE Trans. on Medical Imaging* 13, 725–752 (1994)
14. Quintanilla, D.J., Sanchez, G.M., Gozalez, R.M., Vega, C.A., Andina, D.: Feature extraction using co-ordinate logic filters and artificial neural networks. In: *Proc. of the 7th IEEE International conference on Industrial Informatics, Cardiff, Wales*, pp. 644–649 (2009)
15. Suckling, J., et al.: ‘The Mammographic Image Analysis Society Mammogram Database. In: *Proc. 2nd Int. Workshop Digital Mammography, York, U.K.*, pp. 375–378 (1994)
16. Singh, S., Bovis, K.: An Evaluation of Contrast Enhancement Techniques for Mammographic Breast Masses. *IEEE Transactions on Information Technology in Biomedicine* 9(1), 109–119 (2005)

Analysis of Cryptographically Replay Attacks and Its Mitigation Mechanism

Arun Kumar Singh* and Arun K. Misra

Department of Computer Science and Engineering,
Motilal Nehru National Institute of Technology, Allahabad, Uttar Pradesh, India
singh_arun7@yahoo.com, akm@mnnit.ac.in

Abstract. Replay attack is a typical breach of secured communication between peers that threatens the very design of authentication and key distribution protocols. The designed and proposed protocols are analysed for its strength and weakness and possible vulnerability of replay attack is analysed. Replay attack is type of man-in-middle attack. In this paper, we analyse the replay attack and its countermeasure and propose a new authentication protocol as a solution of replay attack in Junhong Li, 2009 protocol. The session key generated is confirmed by the attacker as a model of secured communication. The sender and receiver without having any idea of such sniffing confidentially enter into the mode of secure communication. All the communication over the session can be hijacked without a trace unless the third party nonce generator is challenged for subsequent renewal and start of new session with the same session key is consciously avoided.

1 Introduction

Encryption system depends largely on the secure system of the key distribution used by the security protocol. Needham and Schroeder (NS) proposed the first important authentication and key distribution protocol in 1978, it happens to be the basis of many authentication protocols [1]. Denning pointed out a flaw of NS protocol in literature in 1981, making people begin to pay attention to the research work in the field of formal security protocol [2]. Replaying attack is a typical issue breach of secured communication between peers that threatens the very design of authentication and key distribution protocols. The generic approach to address such issues is to run it over a standardized security protocol like TLS or SSH, it is sometimes more appropriate to secure the messages directly, i.e., enrich them with security-related parts or wrap them in a secure “container” message or a digital envelop. The messages are not only secured during transport, but they can be stored or passed on including their security features. Our aim of the security goals is that of authenticated message exchange. The server wants to be convinced that the request originated from the alleged client and is not an (unauthorized) duplicate, and the client wants to be convinced that the response originated from the server and is a response to its request. Our proposed models will later allow us to express this formally. The formal analysis

* Corresponding author.

and design is not the part of this paper due to space limitation and will be taken up as a separate study in future.

2 Cryptosystem Attack

A replay attack is a form of network attack in which a valid data transmission is maliciously or fraudulently repeated or delayed [3]. This is carried out either by the originator or by an adversary who intercepts the data and retransmits it, possibly as part of a masquerade attack by IP packet substitution (such as stream cipher attack).

3 Authentication Protocols and Cryptography

In 1976, Diffie and Hellman proposed the idea of public key [4], Rivest, Shamir and Adleman proposed the famous RSA public key algorithm in 1978. Public key cryptosystems do not have to adhere to distribution of the private key system. The distribution of keys will not be required for establishing secure communication. But the efficiency of the algorithm is slower than the private key and is not suitable for large amounts of data encryption [1]. It confirms the identity of both communicating parties and distributes session key, guarantees the confidentiality of information transmitted. In 1983, Dolev and Yao pointed out a protocol can be designed on the assumption that cryptosystem and technology are good in the sense of reliability [1].

4 Analysis of Cryptographic and Authentication Protocols

In various cryptographic scheme, Alice and Bob would like to schedule a meeting time which Alice decides on; she encrypts the time with Bob's key, and sends it to him.

4.1 Concept of Replay Attack

An adversary can intercept Alice's message, and resend it at a later time. In that first phase, the attacker is passive in the sense that he only catches all the messages and eventually keeps some copies that are sent to Bob from Alice. Concept of Replay attack:

Original Protocol:

Alice → *Bob*: {30MAY11,12h10} K_B
Alice → *Bob*: {30MAY11,12h10} K_B

Attack:

Alice → *Ivan*{30MAY11,12h10} K_B → *Bob* {30MAY11,12h10} K_B
Alice → *Ivan*{31MAY11,12h45} K_B → *Bob* {30MAY11,12h10} K_B

Alice chooses a meeting on 30th May, 2011 at 12:10 and send to Bob, Ivan keep this message only in passive mode i.e. listing mode. Suddenly, due to some changed circumstances, Alice changes meeting date and time and sends to Bob but attacker already intercepts the previous message. He can now send the message to Bob again.

So Bob never knows about the updated date and time (31st May-2011, 12:45). This attack is a classic case of Man-in-the-Middle (MITM) attack. It is still valid for all encryption schemes without authentication.

4.2 Diffe-Hellman (D-H) Authentication Protocol

The Diffe-Hellman (D-H) key exchange protocol is a method of exchanging keys over an insecure channel [4]. Alice and Bob agree to choose two public parameters g and p , where p prime number, g generator publicly available, Alice and Bob choose a random number a and b respectively.

$A \rightarrow B: \{g^a\}$

$B \rightarrow A: \{g^b\}$

$A: A \text{ calculates } K_{ab} \{ \{g^b\} \}^a = g^{ab}$

$B: B \text{ calculates } K_{ba} \{ \{g^a\} \}^b = g^{ba}$

Alice and Bob Compute same session key $K = K_{ab} = K_{ba}$

4.3 MITM Attack on Diffe-Hellman

A is sending g^a , attacker generates z and sent g^z to Bob and Bob generates b and sent g^b to Alice but attacker intercepts that message and sends g^z to Alice. Now, Alice compute $K_1 = g^{az}$, Bob compute $K_2 = g^{bz}$. No session key g^{ab} , in this case is generated. However two parallel sessions between Alice and Attacker and Bob and Attacker are possible. This is case of man-in-middle-attack, also known as parallel session attack.

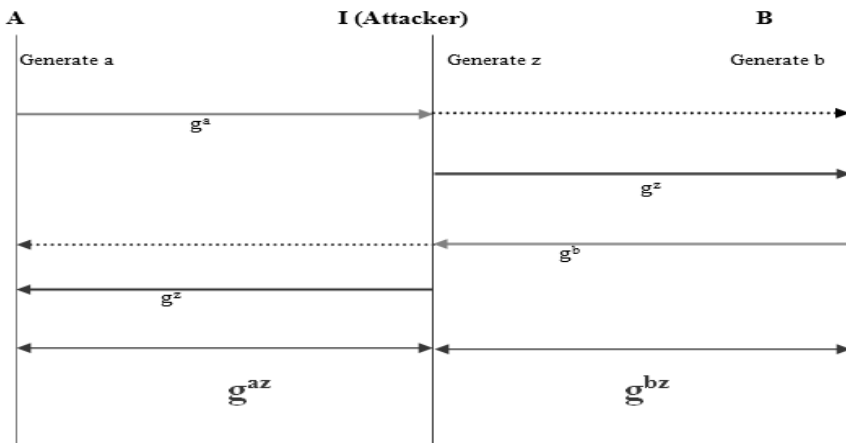


Fig. 1. MITM Attack on Diffe-Hellman

4.4 Analysis of Needham Scroceder (NS) Protocol

Needham Scroceder (NS) protocol was first published in 1987 [6]. It aims to provide mutual authentication between two parties in the initiated session key is a public key

authentication protocol. K_A , K_B are public key for A and B respectively and N_A , N_B are public key and/or nonce generated by A and B respectively.

$$\begin{aligned} A \rightarrow B: & \{N_A, A\}K_B \\ B \rightarrow A: & \{N_A, N_B\}K_A \\ A \rightarrow B: & \{N_B\}K_B \end{aligned}$$

Alice generate a nonce N_A and her identity encrypted by public key of B i.e. K_B , sends to Alice, so this is the initiation of the communication. Bob sends challenge to Alice after the generated nonce N_B and with previous nonce N_A encrypted by public key of A, K_A . Avoidence of replay attack is on nonce generation and its instant use or freshness identifier.

4.5 Lowe Attack

NS protocols is very much secure since 1978, but G. Lowe found its vulnerable in 1995 to a MITM [5]. If the attacker is able to intercept Alice then he can simultaneously relay the message to Bob, Bob thought he received message from Alice.

$$\begin{aligned} A \rightarrow I: & \{N_A, A\}K_I \\ I \rightarrow B: & \{N_A, A\}K_B \\ B \rightarrow I: & \{N_A, N_B\}K_A \\ A \rightarrow I: & \{N_B\}K_I \\ I \rightarrow B: & \{N_B\}K_B \end{aligned}$$

Alice sends a nonce N_A to Ivan using Ivan's public key. Alice is totally aware that she is speaking to Ivan without understanding the security breach. Ivan decrypts the message from Alice and re-encrypts it using Bob's public key. By doing so, he is pretending to Bob that Alice wants to communicate with him. Bob sends to Alice the nonce he has generated and N_A . As the man-in-the-middle, Ivan intercepts the message but he is unable to decrypt it, so he just forwards it to Alice. At this point, Alice thinks that the nonce N_B she has received was actually generated by Ivan, so she sends it back to him using his public key. Ivan is now able to learn the value of N_B , so he can send it to Bob after encrypting it with Bob's public key. When receiving the message, Bob has no evidence that the person with whom he is speaking is not Alice. Therefore, after this attack, Ivan is authenticated as being Alice from Bob's point of view. In fact a partial impersonation takes place. Countermeasure of Lowe Attack:

$$\begin{aligned} A \rightarrow B: & \{N_A, A\}K_B \\ B \rightarrow A: & \{N_A, N_B, B\}K_A \\ A \rightarrow B: & \{N_B\}K_B \end{aligned}$$

4.6 Type Flow Attack on the Needham-Schroeder-Lowe protocol

Type flow is when $A \rightarrow B: M$ and B accepts M as valid but parses it differently. That is that B interprets the bits differently than A. For example, two 16-bit nonces $\{N_A, N_B\}$ could be mistaken as a 32-bit shared key. Type Flow Attack on the Needham-Schroeder Lowe protocol:

$$\begin{aligned} I \rightarrow B: & \{N_I, A\}K_B \\ B \rightarrow I: & \{N_I, N_B, A\}K_A \\ I \rightarrow A: & \{I, (N_B, B)\}K_A \end{aligned}$$

$A \rightarrow I: \{ N_B, B \}$
 $I \rightarrow B: \{ N_B \} K_B$

First of all, Ivan sends to Bob first message of the mutual authentication protocol with the identity of Alice A. Ivan cannot decrypt the message since it is encrypted using Alice's public key, so he just forwards it to Alice. $\{I, (N_B, B)\} K_A$ is the first message Alice receives, and so interprets it as the start of a new protocol run, taking the field $\{N_B, B\}$ to be an agent's identity, and so believes this message came from $\{N_B, B\}$, therefore tries to request $\{N_B, B\}$'s public key, by sending the identity $\{N_B, B\}$ to the server which stores the public keys. Ivan is able to intercept the request for the $\{N_B, B\}$'s public key sent by Alice. Thus, it allows Ivan to learn N_B and so, to respond to the nonce challenge and completed the authentication process between Ivan and Bob, with Ivan who impersonates Alice.

4.7 Authentication Protocol by Li [Junhlong Li, 2009]

A sends identifier A, and the ticket contains A's name, B's name, N_A and randomly generated session key K_{ab} to B. B obtains the ticket from the first step, then B sends the ticket, identifier B and new ticket $\{A, N_b\} K_{bs}$ to key distribution centre S. S receives principal identifier, relating nonce values and session key K_{ab} from the step 2. Then S reorganizes a pair of new tickets and sends them to B. B decrypts message in step 3, receives K_{ab} , and verifies N_b . Then obtains N_A and confirms A's identity. Then generates a new N'_b , encrypts N_a and N'_b by K_{ab} , finally sends it to A. A decrypts message received in step 4 to confirm N_a equal to the N_a in step 1, then send N'_b-I encrypted by K_{ab} to B. Finally B receives message received in step 5 and certificates N'_b-I , then confirms that A really knows session key K_{ab} , thus the protocol achieves the functions of authentication and key distribution.

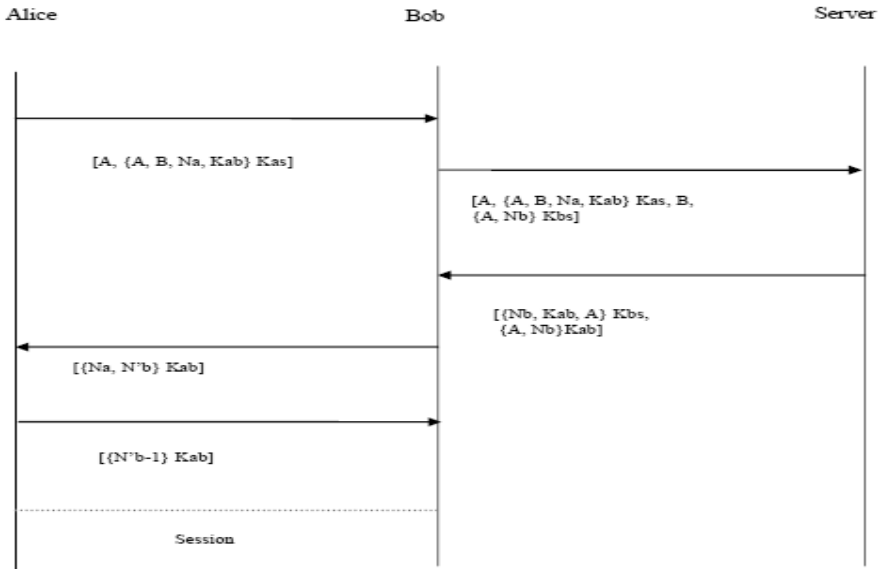


Fig. 2. Key Distribution Protocol [1]

5 Proposed Attack and Its Solution

We specify and analyze protocols for securing authentication and countermeasures of replaying attack. Authentication protocol inspired by web communication security is a system. Notions of authentication as well as the modeling and analysis of cryptographic protocols in different security environment are possible in the analysis.

More precisely, we assume that an existing service or protocol consists of exactly two messages by different parties, request and response or challenge and response, and we specify new protocols that view those messages as part of authentication. We specify concrete and practical protocols for three security goals, namely signature-authenticated two-way authentication, confidential signature-authenticated two-way authentication, and MAC based authenticated two-way authentication as the part of our proposed solution.

5.1 Replay Attack on Li Protocol

Li suggested protocol to mitigate the replay attack in [1], In this protocol attacker (B') intercept the message, when A is sending message to B, So B has $[A, \{A, B, N_a, K_{ab}\}K_{as}]$ then B' sends the message $[A, \{A, B, N_a, K_{ab}\}K_{as}, \{A, N_c\}K_{cs}]$ to Server (S) after that S replays $[\{N_c, A, K_{ab}\}K_{cs} \{A, N_a\}K_{ab}]$ to B. Now B intercepts the K_{ab} and sends message to A $[\{N_a, N_c\}K_{ab}]$.

A replay back to B' $[\{N'_b-1\}K_{ab}]$ has taken place and A thinks he is in communication with B but session has been established as actual communication between A and B' (Attacker) and session has been established between A and B'.

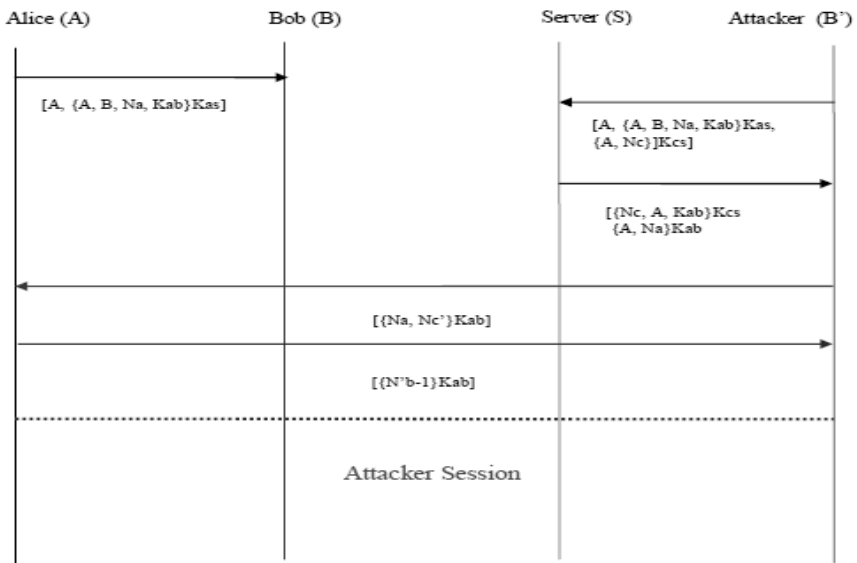


Fig. 3. Proposed Replay Attack on Li Key Distribution Protocol

5.2 Proposed Scheme of Freshness Identifier

When A wants to communicate with B, he sends a request to KGC for B’s public key. Alice or A sends a g^a to Bob by using nonce as well as time stamp matching with the clock synchronization. All the communication from Alice to Bob is totally secure and each stages is secured as authentication is required as well as verification. No chance for replay attack has been left and the design is based on certificate less PKI concept, so key escrow problem has been totally out. Partially dependent on KGC for exchanging the message for authentication happens to be the key of security. If KGS comes under breach of security or compromising the security, then attacker will not able to trace the communication easily. This proposal under various security prospective has been tested by analysis and design by the following set of full proof dialogues.

$A \rightarrow S: \{ID_A, Req\ K_A=?, ReqD_A=?\}K_{AS}$
 $S \rightarrow A: \{A, K_A, D_A\}K_{AS}$
 $A: \{xA\} S_A = x_A, D_A; \{S_A, K_A\}$
 $A \rightarrow S: \{A, ReqB=?\}K_{AS}$
 $S \rightarrow A: \{K_B, B\}K_{AS}$

 $B \rightarrow S: \{ID_B, ReqK_B=?, ReqD_B=?\}K_{BS}$
 $S \rightarrow B: \{B, K_B, D_B\}K_{BS}$
 $B: \{xB\} S_B = x_B, D_B; \{S_B, K_B\}$
 $B \rightarrow S: \{B, ReqA=?\}K_{BS}$
 $S \rightarrow B: \{K_B, A\}K_{BS}$

 $A \rightarrow B: \{\{N_B, a\}K_B\}S_{Awith\ timestamp} \parallel H\{N_B, ID_A, 000ID_B\}\}K_B$
 $B \rightarrow A: \{Ack\ Seq=seqnum_B, H\{N_B, N_A, c\}\}K_A$
 $A \rightarrow B: \{ID_A, seqnum_B, Timestamp\}K_B$
 $B \rightarrow A: \{H\{N_A, b\} \parallel Signature_{Timestamp}\}K_A, H\{a*b\}\}K_A$

 $Session\ Key\ K = K_{AB} = K_{BA} = H\{\{g^a\} \parallel \{g^b\} \parallel \{g^{ab}\} \parallel \{ID_A\} \parallel \{ID_B\} \parallel \{K_A\} \parallel \{K_B\}\}$

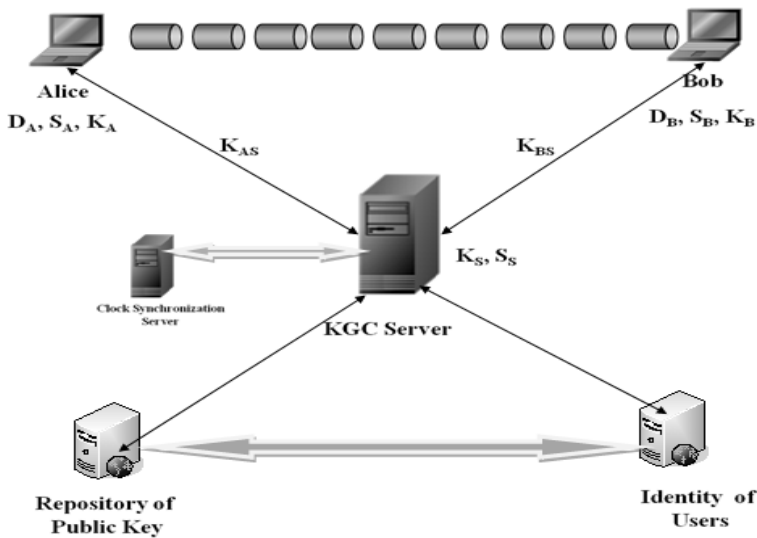


Fig. 4. Secure Communications by Authentication Server

6 Conclusion

Any authentication or key distribution protocol faces the vulnerability due to replay attack. Replay attack has been tackled earlier by specifically designing authentication protocol to prevent such attacks. In this process, there are some protocols designed that expose themselves also to such vulnerability. This paper provides an overview of such attacks and its countermeasure. The earlier designed protocol, the loop holes are analysed. It has been shown successfully that the protocol designed by Li has such a loop hole. This solution to plug the loop hole has been proposed with dialogue generation. The dialogue generation makes use of message freshness based on the freshness identifier. The principle to use freshness identifier has been followed to prevent the flaws in Li Authentication and key distribution protocol. The study can be undertaken to devise and use freshness identifiers to other kinds of MITM attacks in future.

References

1. Li, J.: College of Mathematics and Information Science Hebei Normal University, Hijiazhuang, 050016 China Ljhxwfl@126.com, Design of Authentication Protocols Preventing Replay Attacks, IEEE (2009)
2. Li, G.: Variations on the Themes of Message Freshness and Replay or the Difficulty in Devising Formal Methods to Analyze Cryptographic Protocols, SRI International Computer Science Laboratory Ravenswood Avenue Menlo Park, California 94025 U.S.A.
3. Wikipedia, <http://www.wikipedia.org>
4. Diffie, W., Hellman, M.E.: New Directions in Cryptography. IEEE Transactions on Information Theory IT-11, 644–654 (1976)
5. Gong, L.: Verifiable-text Attacks in Cryptographic Protocols. In: Proceedings of IEEE INFOCOM 1990, pp. 686–693 (1990)
6. Needham, R., Schroeder, M.: Using Encryption for Authentication in Large Networks of Computers. Communications of the ACM 21(12) (December 1978)
7. Otway, D., Rees, O.: Efficient and Timely Mutual Authentication. Operating System Review 21(1) (January 1978)
8. Singh, A.K., Tewari, P., Samaddar, S.G., Misra, A.K.: Communication Based Vulnerabilities and Script based Solvabilities. In: International Conference on Communication, Computing & Security (Proceedings by ACM with ISBN-978-1-4503-0464-1), February 12-14, National Institute of Technology Rourkela Orissa, India (2011)
9. Singh, A.K., Tewari, P., Samaddar, S.G., Misra, A.K.: Vulnerabilities of Electronics Communication: solution mechanism through script. International Journal of Computer Science Issues (IJCSI) 8(3) (2011)

Comparisons of Three Classifier for Classification of Bamboo Plant

Krishna Singh^{1,*} and Surendra Singh²

¹ Department of Electrical Engineering,
Indian Institute of Technology Roorkee,
Roorkee, 247667

singhkrishna5@gmail.com

² Bharat Sanchar Nigam Limited
dgmsurendra@gmail.com

Abstract. Tropical rainforest has more than 3,000 different types of timber species, out of these, about 200 species are being used by the timber industry. The properties of these bamboo species varies a lot and different species are recommended for different purpose. Due to this fact, recognition of bamboo species is necessary before its efficient utilization. Due to unavailability of the database, the database is developed in *Forest Research institute Dehradun* by collecting the raw samples of Culm sheath. The three moment based classification techniques i.e. Central moment Legendre moment and Fourier moment is adopted to perform the experiment. The performance of these techniques is measured by introducing three parameters i.e. classwise classification accuracy, overall classifier accuracy and computation time. A confusion matrix is created to quantify the class wise and classifier accuracy. The results show that the Fourier Moment has to be superior classification accuracy compared to Legendre and central moment, and computation time is very low, because only boundary points are considered for calculating the moment.

This application can eliminate the need for laborious human recognition method requiring a plant taxonomist. The results obtained shows considerable recognition accuracy proving that the techniques used is suitable to be implemented for commercial purposes.

Keywords: Culm Sheath, confusion matrix, precision and Recall.

1 Introduction

Bamboo has also a long and well established tradition for being used as a construction material throughout the tropical and sub-tropical regions of the world from a long time.[16]. In the modern context when forest cover is fast depleting and availability of wood is increasingly becoming scarce, the research and development undertaken in past few decades have established and amply demonstrated that bamboo could be a viable substitute of wood and several other traditional materials for housing and building construction sector and several infrastructure works. Its use through industrial processing has shown a high potential for production of composite materials and components which are cost-effective and can be successfully utilized for structural and nonstructural applications, accordingly more plantation or production

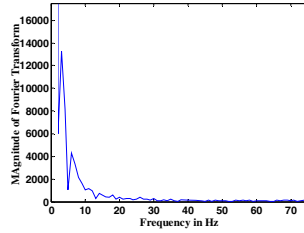
of good quality bamboo is to be planned in near future. Due to this fact, successful application of bamboo in engineering relies on the selection of a correct species. In order to assess the value of particular bamboo species it is necessary first to be able to identify them accurately [3, 4, 5]. The identification can be done mainly by a systematic botanist with a long experience in this field. In traditional taxonomy, components of flowers are often used as the most important part of the plant to classify a species, but in some bamboo species flowering interval can be up to 130 years [3, 4]. The shapes of bamboo Culm sheath provide valuable data for identification of bamboo species and are readily available. A central management database system would be developed that shows which bamboo species is used for which purpose. By image processing and the three moment based classifiers viz central moment, Legendre moment and Fourier moment classifier the proposed work is for classifying bamboo of five different species, using the shape features of Culm sheath (modifying Leaf of bamboo) [1, 2, 9, 10, 12, 18]. In pattern recognition application, there are lot of work reported by the researcher for retrieval and classification of various object shapes by the generic Fourier descriptor, Legendre moment and wavelet Zernike moment etc [6, 8, 12, 15, 18, 19, 20, 21]. A set of moment invariants using a nonlinear combination based on normalized central moments for shape recognition is introduced [9]. Some recent work focused on recognition and classification of plant by their leaves features. The plant classification by their leaves using Image moment based classifier for calculating the shape moments of leaves, and compares the performance of classifiers [10]. The classification of plant species by Image processing and neural network techniques is reported [11-14] using leaves geometrical and morphological features. The taxonomy of bamboo is introduced by [3,4,5]

1.1 Methodology for Developing Database

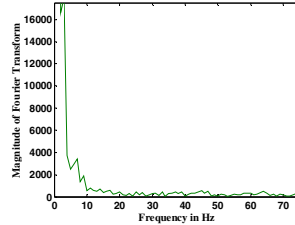
Culm sheath of five different species of bamboos viz. *Bambusa vulgaris*, *Bambusa Balcooa*, *Bambusa Tulda*, *Dendrocalamus Melingensis*, *Dendrocalamus Longipathus* are collected from the bambusetum of *Forest Research Institute, Dehradun*. Culm sheaths at the Culm base are different from those higher up. They are broader and have shorter blades. [12]. Collection of Culm sheaths in good condition is important for further studies on automation. Culm sheath from bamboo plant is firstly removed with a sharp knife or blade because it is very hard when it is green. The sheaths are then cleaned to remove hairs from the surface of the sheath. The same is then pressed with a heavy weight of plane object to flat the Culm sheath. In whole procedure some Culm sheath got broken while removing/cleaning the hair, some developed cracks in vertical direction; therefore problem to scan the actual boundary of the samples Culm sheath is faced.

Since the image database of the Culm sheath is not available from any other source therefore the samples of Culm sheath are collected from *Forest research Institute Dehradun* to determine both intraspecific and interspecific variation. For acquiring the image of Culm sheath, two procedures are followed to capture the image of sheath, firstly the smaller ones through scanner and secondly the bigger ones through camera wizard. For acquired the image through camera ,it is first laid on a white sheet and then captured the image using a high resolution camera from a distance of approximately 3 feet from the ground level. The acquired image are not very much clear, because of some of the boundary part are broken while removing the hair on the

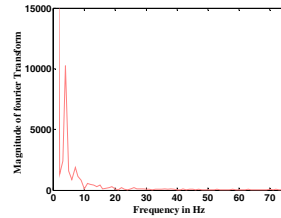
surface of the Culm sheath, Therefore image pre-processing is required to enhance the quality of the acquired image. By image processing colored image are converted into gray level then in binary images using segmentation and thresholding [7, 17] method and the output image is a binary image in which the leaf object are numerically displayed with 1 and the background is with 0. The sample image of the Five species and their Fourier representation is shown in Fig. 1



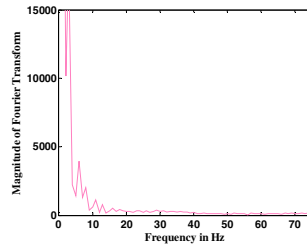
(a)



(b)



(c)



(d)

Fig. 1. Sample images of culm shaeth and their Fourier representation(a) *Bambusa vulgaris* (b) *Bambusa Balcooa* (c), *Bambusa Tulda* (d) *dendrocalamus Melingensis* (e) *Dendrocalamus Longipathus*

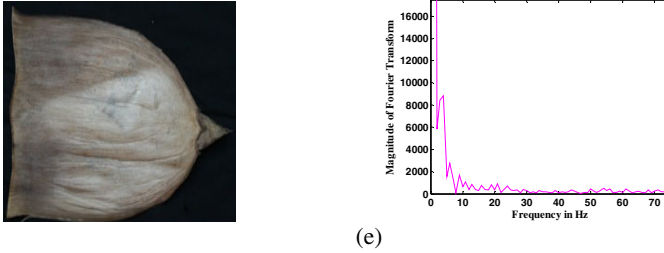


Fig. 1. (continued)

2 Classification Techniques

The Three moment based classification techniques viz Central moment, Legendre moment and Fourier moment are applied on the dataset of culm sheath of bamboo species.[10] The performance of the classifier is measured by two parameters, the Overall classifiers accuracy (OCA) and Class wise classification accuracy (CWCA). The confusion matrix shows the class wise accuracy of each class. The diagonal elements of all the classes correspond to correct classification while all off diagonal elements indicate classification errors. The effectiveness of a classifier is determined by summing all the diagonal elements and dividing this sum by the sum of all elements of the matrix. This metric measured the overall classifier accuracy. Both OCA & CWCA are calculated using confusion matrix. The average classification accuracy of the three classifiers namely central moment, Legendre moment and Fourier moment are observed. Computational complexity is a crucial point in all aspects of image and pattern recognition. The computational complexity of all the three moment based classifier are also measured to evaluate the performance of the classifier

3 Experimental Results

The experiments are carried out on five species of bamboo Culm sheath viz *Bambusa Vulgaris*, *Bambusa Balcoa*, *Bambussa Tulda*, *Dendrocalamus membranaceus* and *Dendrocalamus Longipathus*. The effectiveness of the classifier is implemented on MATLAB 7.6 on an Intel Core 2 Quad (processor) computer, Windows Vista operating system with 1.83 GHz CPU and 3 GB RAM. The dataset of bamboo Culm sheath has total 46 samples. *Bambusa Vulgaris* 17, *Bambusa Balcoa* 9 and *Bambussa Tulda* 9. *Dendrocalamus Melingensis* 6 and *Dendrocalamus Longipathus* have 5 samples.

The accuracy of the each classifier is calculated by designing a confusion matrix shown in table 1, 2 and 3 for central moment, Legendre moment and Fourier moment respectively. The overall performance of the three classifiers w. r. t accuracy and

computation time is shown in table 4. It is observed that the overall classification accuracy of the central moment is 41.06 % Legendre moment is 55.99% and Fourier moment is 75.56. It is observed that Fourier moment saved 99.64 % as compare to Legendre moment and 95.99 as compared to central moment, since it is computed from a 1-D function that represents the shape boundary points and the geometric invariance is also achieved after the Fourier transform by normalizing Fourier coefficients. The Legendre moment is more computationally expensive because much computation is required to calculate the complex quantity.

It is concluded that the Fourier moment classifier has better performance compared to Legendre moment and central moment classifier. The Legendre moment is computationally expensive as compared to central moment, because boundary as well as region point are considered to calculate the moment, but accuracy is better than central moment because it has less information redundancy. The graphical representation of the accuracy of the three classifier is shown in Fig. 2.

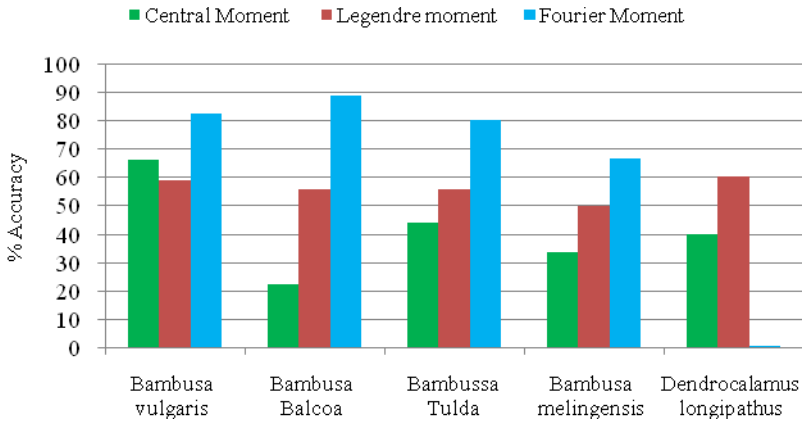


Fig. 2. Accuracy comparisons of the three classifiers

Table 1. Confusion Matrix in Central Moment

# species Name	<i>Bambusa vulgaris</i>	<i>Bambusa Balcoa</i>	<i>Bambussa Tulda</i>	<i>Bambusa Melingensis</i>	<i>Dendrocalamus Longipathus</i>
<i>Bambusa vulgaris</i>	11	0	3	0	3
<i>Bambusa Balcoa</i>	3	2	0	4	0
Bambussa Tulda	3	0	4	0	2
<i>Bambusa melingensis</i>	1	3	0	2	0
<i>Dendrocalamus longipathus</i>	1	0	0	2	2

Table 2. Confusion Matrix in Legendre Moment

# species Name	<i>Bambusa vulgaris</i>	<i>Bambusa Balcoa</i>	<i>Bambusa Tulda</i>	<i>Bambusa Melingensis</i>	<i>Dendrocalamus Longipathus</i>
<i>Bambusa vulgaris</i>	10	0	5	0	2
<i>Bambusa Balcoa</i>	0	5	0	4	0
<i>Bambusa Tulda</i>	3	0	5	0	1
<i>Bambusa melingensis</i>	0	3	0	3	0
<i>Dendrocalamus longipathus</i>	2	0	0	0	3

Table 3. Confusion Matrix of Fourier Moment

# species Name	<i>Bambusa vulgaris</i>	<i>Bambusa Balcoa</i>	<i>Bambusa Tulda</i>	<i>Bambusa Melingensis</i>	<i>Dendrocalamus Longipathus</i>
<i>Bambusa vulgaris</i>	14	0	2	0	1
<i>Bambusa Balcoa</i>	0	8	0	1	0
<i>Bambusa Tulda</i>	0	0	8	0	1
<i>Bambusa melingensis</i>	0	1	1	4	0
<i>Dendrocalamus longipathus</i>	2	0	0	0	3

Table 4. Comparison of Overall Performance of the three classifiers

# Bamboo species	Accuracy (%)			Computation Time (seconds)		
	Fourier Moment	Legendre Moment	Central Moment	Fourier Moment	Legendre Moment	Central Moment
<i>Bambusa Vulgaris</i>	82.3	58.82	66.0	.0185	4.32	0.213087
<i>Bambusa Balcoa</i>	88.8	55.55	22.0	.0175	2.73	0.440303
<i>Bambusa Tulda</i>	80	55.56	44	.0034	4.12	0.327964
<i>Dendrocalamus Melingensis</i>	66.7	50.0	33.33	.0190	2.89	0.41442
<i>Dendrocalamus longipathus</i>	60.0%	60.0	40	.0046	3.74	0.26120
% Average Performance of the classifiers	75.56%	55.99%	41.06%	0.0126	3.556	0.33068

4 Conclusion and Scope for Future

As a shape descriptor technique, the evidence to date is that Fourier moment are very good features to use when dealing with particular types of shapes. The aim of the

proposed work is to investigate the usefulness of Fourier descriptors for the shape description for bamboo culm sheath. Thus, based on the above results it can be concluded that Fourier moment is one of the several techniques of object recognition that produces optimal results for bamboo species classification.

References

1. Imaya, A.: Fourier analysis of Three-Dimensional Shapes. In: SPIE Conference on Vision Geometry VII, vol. 3454, pp. 87–98 (1998)
2. Reeves, A.P., Prokop, R.J., Andrews, S.E., Kuhl, F.: Three-dimensional shape analysis using moments and Fourier descriptors. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 10(6), 937–943 (1988)
3. Bhattacharya, S., Das, M., Bar, R., Pal, A.: Morphological and Molecular Characterization of *Bambusa tulda* with a note on Flowering, pp. 529–535. Published by Oxford University Press on behalf of the *Annals of Botany* 98 (2006)
4. Stapleton, C.: Bamboo of Nepal: An illustrated guide published by forestry research and information centre. In: Department of forestry and plant research, Government of Nepal Kathmandu
5. Rodrigues, C.S., Gomes, O.D.M., Ghavami, K., Paciornik, S.: A classification system of bamboo based on the digital image processing of its Mesostructure. In: 11th International Conference on Advance Material, Rio De janerio Brazil, September 20-25, (2009)
6. Persoon, E., Fu, K.S.: Shape Discrimination Using Fourier Descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 8, 388–397 (1986)
7. Gonzalez, R.C., William, P.: *Digital Image Processing*. Addison-Wesley (1977)
8. Granlund, G.H.: Fourier Pre-processing for Hand Print Character Recognition. *IEEE Transactions on Computer* 21, 195–201 (1972)
9. Hu, M.K.: Visual pattern recognition by moment invariants. *IEEE Transaction on Information Theory* 8, 179–187 (1962)
10. Singh, K., Gupta, I., Gupta, S.: Classification of leaf image using Image moment based Classifier. In: *Information Technologies and Energy Management*. Excel India Publisher, New Delhi (2010) (ISBN:978-93-80697-07-9)
11. Singh, K., Gupta, I., Gupta, S.: Plant Species Classification By leaves using neural network. In: *International conference on trends and Advancement in Electronics and Computer*, February 25-26, pp. 51–58 (2010)
12. Singh, K., Gupta, I., Gupta, S.: Classification of Bamboo Species by Fourier and Legendre Moment. *International Journal of advanced Science and Technology* (accepted)
13. Singh, K., Gupta, I., Gupta, S.: SVM-BDT PNN and Fourier Moment Technique for Classification of Leaf Shape *International Journal of Signal Processing, Image Processing and Pattern Recognition* 3(4) (December 2010)
14. Singh, K., Gupta, I., Gupta, S.: Comparison of PNN-PCA with SVM-BDT and Moment based technique for leaf shape recognition and classification. In: *International Conference on Image Processing Computer Vision, & Pattern Recognition*, Las Vegas, USA, July 12-15 (2010) ISBN:1-60132-152-X,1-60132-153-8,1-60132-154-6 printed at USA 2010
15. Keyes, L., Winstanley, A.C.: Fourier Descriptors as a General Classification Tool for Topographic Shapes. In: *IMVIP Proceedings of the Irish Machine Vision and Image Processing Conference*, Dublin City University, pp. 193–203 (1999)

16. Ministry of Urban Development & Poverty Alleviation Government of India, Bamboo: A material for cost effective and disaster resistant housing Building Materials and Technology Promotion Council
17. Otsu, N.: A threshold selection method from gray level histograms. *IEEE-T SMC* 9(1), 62–79 (1979)
18. Mukundan, R., Ramakrishna, K.R.: Fast computation of Legendre and Zernike moments. *Journal of Pattern Recognitions* 28(9), 1433–1442 (1995)
19. Yadav, R.B., Nishchal, N.K., Gupta, A.K., Rastogi, V.K.: Retrieval and classification of objects using generic Fourier, Legendre moment, and wavelet Zernike moment descriptors and recognition using joint transform correlator. *Journal of Optics & Laser Technology* 40, 517–527 (2008)
20. Pakchala, S., Lee, P.: Pattern recognition in gray level images using moment based invariant features. In: *IEEE Conference Publication on Image Processing and its Applications*, vol. 465, pp. 245–249 (1999)
21. Teague, M.R.: Image analysis via general theory of moments. *Journal of optical society of America* 70, 920–930 (1980)

Secure GKA Using SVD Matrix Decomposition and Kronecker Product

Reddi Siva Ranjani¹, D. Lalitha Bhaskar², and P.S. Avadhani³

Department of CS&SE, Andhra University, Visakhapatnam
{rsivaranjani_55, psavadhani}@yahoo.com
lalithabhaskari@yahoo.co.in

Abstract. Today many applications require group communication. To deploy cooperation among the members, secure multicast service must be provided efficiently and safely exchange data among the group members. A common key is to be used by the group members for safe and secure communication. Diffie–Hellman [5] is the key agreement scheme used for sharing the common key by using the public channels. This paper describes a new method for secure method for Group Key Agreement (GKA) using SVD matrix decomposition and kronecker product. SVD matrix decomposition is used for factorizing the matrix and kronecker product is used for computing the common key.

Keywords: Group Management, Key Agreement, kronecker product, SVD Matrix.

1 Introduction

Group key management is the main issue in secure group communication [4]. Group communication is the main theme; members in the group are dynamic. Whenever an existing member leaves or new member joins, a new group key is computed with the help of group members. Group key agreement protocol (GKAP) is one of the basic cryptographic protocols. GKAP allows two or more parties negotiate a common secret key using insecure communications. First key agreement protocol was presented by Diffie-Hellman, which caused rapid development of asymmetric cryptography, allowing two users communicating over a public insecure channel to agree on a common shared secret key. This paper is using the SVD matrix decomposition concept for securely transferring the part of key to the other group members. SVD matrix decomposition is used for factorizing the matrix into three parts. In this paper fragment is transferred onto other member, the other member will send his fragment to the sender; this will avoid the third party to get the actual matrix in key computation. Once the fragments are completely transferred between each other, the common key is computed using kronecker product.

The rest of the paper is organized as follows. Section 2 describes the mathematical background of SVD matrix and kronecker product used in the group key agreement. Section 3 gives the key agreement algorithm Section 4 presents the experimental results using MATLAB and Finally, Section 5 concludes this paper.

2 Mathematical Background of SVD

The Singular Value Decomposition (SVD) [1,2] is a widely used technique to decompose a matrix into several component matrices, exposing many of the useful and interesting properties of the original matrix. The decomposition of a matrix is often called a factorization. Ideally, the matrix is decomposed into a set of factors (often orthogonal or independent) that are optimal based on some criterion. Using the SVD, one can determine the dimension of the matrix range or more-often called the rank. The rank of a matrix is equal to the number of linear independent rows or columns. This is often referred to as a minimum spanning set or simply a basis. The SVD can also quantify the sensitivity of a linear system to numerical error or obtain a matrix inverse. Additionally, it provides solutions to least-squares problems and handles situations when matrices are either singular or numerically very close to singular.

SVD decomposes a regular matrix $A_{n \times n}$ into three matrices U, S and V^T . The formula for getting the matrix A is given by

$$A_{n \times n} = U_{n \times n} X S_{n \times n} X V_{n \times n}^T$$

It also decomposes a regular matrix $A_{m \times n}$ into three matrices U, S and V^T . The formula for getting the matrix A is given by

$$A_{m \times n} = U_{m \times m} X S_{m \times n} X V_{n \times n}^T$$

Where S =diagonal matrix U =left eigenvector matrix V^T =right eigenvector matrix

Procedure for Computing $U_{m \times m}$ and $V_{n \times n}$ matrix: The matrix $U_{m \times m}$ and $V_{n \times n}$ can be computed by using eigenvalues. Compute the eigenvalues compute in exactly the same manner the eigenvectors of AA^T and $A^T A$. Once these are computed we place these along the columns of U and V respectively.

Procedure for computing $S_{m \times n}$ matrix: The $S_{m \times n}$ isa diagonal matrix consisting of r non-zero values in descending order. The steps for computing S matrix as:

Step1: A^T and $A^T A$ were computed.

Step 2: the eigen values of $A^T A$ were determined and sorted in descending order, in the absolute sense. The nonnegative square roots of these are the singular values of A .

Step3: S was constructed by placing singular values in descending order along its diagonal.

Following special properties of SVD decomposition is used in the key agreement.

Property 1: Suppose A is $m \times n$ matrix, if $m \geq n$ then the inverse of A exists, calculated by

$$A^{-1} = (A^T A)^{-1} A^T, \text{ then } A^{-1} A = I$$

Property 2: Suppose A is $m \times n$ matrix, if $m \leq n$ then the inverse of A exists, calculated by

$$A^{-1} = A^T (A^T A)^{-1}, \text{ then } A A^{-1} = I$$

Kronecker Products: The Kronecker product[3] of two matrices A and B of sizes $P \times Q$ and $R \times S$, respectively, is defined as

$$A = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \end{bmatrix}, B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \text{ then } A \otimes B = \begin{bmatrix} A_{11}B & A_{12}B & A_{13}B \\ A_{21}B & A_{22}B & A_{23}B \end{bmatrix}$$

$$= \begin{bmatrix} A_{11}B_{11} & A_{11}B_{12} & A_{12}B_{11} & A_{12}B_{12} & A_{13}B_{11} & A_{13}B_{12} \\ A_{11}B_{21} & A_{11}B_{22} & A_{12}B_{21} & A_{12}B_{22} & A_{13}B_{21} & A_{13}B_{22} \\ A_{21}B_{11} & A_{21}B_{12} & A_{22}B_{11} & A_{22}B_{12} & A_{23}B_{11} & A_{23}B_{12} \\ A_{21}B_{21} & A_{21}B_{22} & A_{22}B_{21} & A_{22}B_{22} & A_{23}B_{21} & A_{23}B_{22} \end{bmatrix}$$

Thus, $A \otimes B$ is a matrix of size $PR \times QS$.

3 Key Agreement Scheme Using SVD

In this Section, we will give a key agreement scheme which is an analogue of the Diffie-Hellman key agreement process by using the SVD decomposition of matrices over a finite field. Part of the group key is shared securely[4] between the group members.

We assume that a user, Alice, wants to establish a common secret key with Bob via a public channel for further secret communications.

They can follow the steps below:

1. Alice randomly chooses a $m \times n$ matrix A and decompose the matrix by using SVD decomposition
2. Alice sends Bob $S_A * V_A^T$
3. Bob randomly chooses a $n \times m$ matrix B and decompose the matrix by using SVD decomposition
4. Bob sends $S_A * V_A^T * U_B * S_B$ to Alice
5. Alice sends $U_A * S_A * V_A^T * U_B * S_B$ to Bob.
6. Bob sends $U_A * S_A * V_A^T * U_B * S_B * V_B^T$ to Alice
7. Alice computes the common shared key $K = A \otimes B$ by using $A \otimes (A^T * A)^{-1} * A^T * U_A * S_A * V_A^T * U_B * S_B * V_B^T$
8. Bob computes the common shared key $K = A \otimes B$ by using $U_A * S_A * V_A^T * U_B * S_B * V_B^T * B^T * (B^T * B)^{-1} \otimes B$

4 Experimental Results Using MATLAB

Algorithm:

- 1: Alice selects a random matrix
- 2: Bob selects a random matrix
- 3: Alice divides the matrix into three parts using SVD decomposition of matrix
- 4: Bob divides the matrix into three parts using SVD decomposition of matrix
- 5: Alice sends part of the decomposed matrix
- 6: Bob sends part of the decomposed matrix i.e. multiplied with received part
- 7: Steps 5 and 6 repeated for three times
- 8: Alice and Bob individually computes the common key

Output:

```

MATLAB Command Window
File Edit View Window Help
>> svd2
Enter A[1,2;3,4]
Enter B[3,4;5,6]
Sender Matrix
     1     2
     3     4

Receiver Matrix
     3     4
     5     6

Senders Key is
  3.0000  4.0000  6.0000  8.0000
  5.0000  6.0000  18.0000 12.0000
  9.0000 12.0000  12.0000 16.0000
 15.0000 18.0000  20.0000 24.0000

Receivers Key is
  3.0000  4.0000  6.0000  8.0000
  5.0000  6.0000  18.0000 12.0000
  9.0000 12.0000  12.0000 16.0000
 15.0000 18.0000  20.0000 24.0000
  
```

5 Conclusion

We study the problem in group key management in mathematical point of view. A simple and secure group key agreement scheme is proposed based on matrix operations. The presented algorithm is for group key agreement for generating a common key by more than one people in the group for secure group communication. Strong evidence has been supplied for the practical implementations of this agreement. This approach is secure, and its backward and forward secrecy can be guaranteed. The security of our approach relies on the fact that user cannot compute the correct inner product without knowing all matrix factors, therefore cannot derive the group key. The advantages of our scheme include: 1) it is not necessary to invoke strong encryption algorithm, the re-keying messages can be broadcast or multicast via open channel, when members register to form the group or new members join in; 2) very efficient and scalable for large size group, 3) can handle massive membership change efficiently; and 4) the computation overhead and storage capacity of group member are both small, which will not increase as the group size grows.

References

1. Conte, S.D., de Boor, C.: *Elementary Numerical Analysis: An Algorithmic Approach*, 3rd edn. McGraw-Hill (1980)
2. Golub, G.H., Van Loan, C.F.: *Matrix Computations*. Johns Hopkins University Press (1983)
3. Laub, A.J.: *Matrix Analysis for Scientists and Engineers*. SIAM: Society for Industrial and Applied Mathematics press (2011)
4. Bohli, J.M., Gonzalez Vasco, M.I., Steinwandt, R.: Secure group key establishment revisited. *Int. J. Inf. Sec.* 6(4), 243–254 (2007)
5. Bresson, E., Chevassut, O., Pointcheval, D., Quisquater, J.J.: Provably authenticated group Diffie-Hellman key exchange. In: *CCS 2001: Proceedings of the 8th ACM conference on Computer and Communications Security*, pp. 255–264. ACM (2001)
6. Wang, E.K., Ye, Y.: An Efficient and Secure Key Establishment Scheme for Wireless Sensor Network. In: *Intelligent Information Technology and Security Informatics (IITSI)*, pp. 511–516 (2010)
7. Dai, H., Xu, H.: Key Predistribution Approach in Wireless Sensor Networks Using LU Matrix. *IEEE Sensors Journal* (8), 1399–1409 (2010)

Time-Frequency Domain Techniques for Power System Transients Identification

Srikanth Pullabhatla¹ and Prasad Chintakayala²

¹ PGET, EDRC-KKR, MMH-IC, Larsen & Toubro Ltd., DLF,
Kolkata, West Bengal, India

² Dept. of Electrical Engineering, Raghu Engineering College,
Dakamarri, Visakhapatnam Distt., Andhra Pradesh, India
{srikanth.srikki, prasaddude88}@gmail.com

Abstract. In the present paper power system transient disturbances using Stockwell and Hartley Stockwell transform have been investigated. It is well known that Stockwell transform (ST) is utilized for the real time prediction of the disturbance as it is able to accurately determine the sudden burst in the signal. However, in some cases the frequency resolution of ST is low. Consequently Hartley Stockwell transform (HST) has been proposed which is good in frequency resolution. A comparison has been made with the proposed HST with ST technique. The proposed HST technique has been tested on various transient disturbances to show its efficacy.

1 Introduction

Due to an increase in the size and power levels, power system has become more complex at present. Consequently, many factors like power quality, transient, instability etc need to be addressed, as these problems influence the behaviour of the power system. With the incorporation of a large number of sensitive and critical loads into the system and owing to the inclusion of deregulation and competition in the power market, utilities are now more concerned in identifying, measuring and monitoring of above mentioned factors. Also necessary corrective actions for their reduction and elimination have become essential. Among a numerous disturbances, a transient disturbance is the outward manifestation of a sudden change in circuit conditions. When a switch opens or closes or a fault occurs on the system such an operating condition is of major concern [1]. Though the operating time of a transient condition is very small compared with the steady state time, the effect of these on the power system is very significant. The major reason for a transient is the presence of inductors and capacitors in the system. A transient is primarily caused by capacitor switching, dynamic load switching, circuit breaker operations etc which are unavoidable for proper operation of the power system. The study of transient periods is extremely important because under such periods, the circuit components are subjected to greatest stresses due to excessive currents or voltages. These high voltages and currents cause breakdown of insulations, damage to windings, inaccurate operation and damage of sensitive loads. Therefore, it is important to identify and mitigate such transient events before these obstruct the normal operation of the system.

Identification of the typical transients problems have been dealt by various researchers using Fourier, Hilbert & Wavelet transforms [2-6]. However, it has been reported in the literature that these methods cannot properly predict the transient disturbance parameters whenever there is a sudden burst in the signal. Keeping in view the concern of the sudden or transient behavior of the signal in power system, there is still a need for some effective tool for predicting the typical disturbances namely capacitor switching, dynamic load switching, inrush currents etc. [7, 8].

Thus the present work is to analyze the transient behaviour using Time Frequency Resolution (TFR) which has better prediction properties. The method employed to predict a transient problem is the modified version of Stockwell transform and is known as Hartley Stockwell Transform (HST) [9]. In Stockwell transform (ST), Gaussian window is used. The technique has a better representation in time-frequency domain [10, 11]. However, compared to HST the frequency resolution is low. HST has not been used for identification of power system transients and the present work is a step in that direction.

The technique utilized to identify a transient disturbance in the time frequency analysis based on ST and HST, is explained in detail in section 2 and section 3 respectively. Different transients which commonly occur in power system are analyzed and presented in Section 4 along with the implications of the results. Finally conclusions are drawn in section 5.

2 Modified Wavelet Transform-S Transform

According to the transformation theory, information in any signal is contained in the phase of the spectrum and its amplitude. S-transform, like other transforms uses this theory to analyze a given signal and is an improved version of Continuous Wavelet Transform (CWT) where the amplitude and phase of the spectrum is converted into information in CWT domain. In order to make use of the information contained in the phase of the CWT, it is necessary to modify the phase of the mother wavelet. The CWT of a function is defined as [10, 11] eqn. 1,

$$W(\tau, \alpha) = \int_{-\infty}^{\infty} h(t) * w(t-\tau, \alpha) dt \quad (1)$$

where, W is a scaled replica of the fundamental mother wavelet, the dilation determines the width of the wavelet and this controls the resolution. The S-transform is obtained by multiplying the CWT with a phase factor as

$$S(\tau, f) = \exp(i2\pi f \tau) * W(\tau, \alpha) \quad (2)$$

where, the mother wavelet for this particular case is defined as

$$W(t, f) = \left(\frac{|f|}{\sqrt{2\pi}}\right) * \exp\left(-\frac{t^2 f^2}{2}\right) * \exp(-i2\pi f t) \quad (3)$$

In eqn. (3), the dilation factor is the inverse of the frequency. Thus, the final form of the continuous S transform (CST) is obtained as

$$S(\tau, f) = \int_{-\infty}^{+\infty} h(t) \left(\frac{|f|}{\sqrt{2\pi}} \right) e^{-\frac{(\tau-t)^2 f^2}{2}} e^{-i2\pi ft} dt \quad (4)$$

and the width of the Gaussian window is

$$\sigma = \frac{\sqrt{1}}{|f|} \quad (5)$$

The linear property of the S transform ensures that for the case of additive noise, one can model the data as:

$$\text{data}(t) = \text{signal}(t) + \text{noise}(t), \quad (6)$$

the S transform gives

$$S\{\text{data}\} = S\{\text{signal}\} + S\{\text{noise}\} \quad (7)$$

2.1 Discrete S-Transform

Since Discrete S-transform (DST) is a representation of the local spectra, it can be obtained by the shift operation on Fourier spectrum and is expressed as eqn. (8),

$$S\left[kT, \frac{n}{NT}\right] = \sum_{k=0}^{N-1} H\left[\frac{m+n}{NT}\right] e^{-\frac{2\pi^2 m^2}{n^2}} e^{\frac{i2\pi mk}{N}} \quad \text{for } n \neq 0 \quad (8)$$

where, k, m vary from 0 to $N-1$, n varies from 0 to $(\frac{N}{2}-1)$ and T = sampling time.

In this paper, the ST amplitude matrix is used to analyze the current waveforms of ac traction in which the rows are the frequencies and the columns are the time values. Each row displays the ST amplitude with all frequencies at the same time and each column displays the ST amplitude with time varying from 0 to $N-1$ in the same frequency. The features necessary for power system transients identification are extracted from the S-matrix. Further, from the S-matrix important information in terms of amplitude, frequency and phase are extracted.

In Fig. 1(a) a chirp function with a high frequency at 20th sample, low frequency at 40th sample and medium frequency from 60th sample is shown. In this signal, there is a sudden change or burst at 20th sample. Fig. 1(b) depicts the contours or S plot of the chirp signal with time (samples) on x-axis and frequency on y-axis. The contours are seen from lower range of values to higher and contain separated contours in all the low, medium and high frequency region as there are high instantaneous and steady changes in the frequency. Contour plot of the time series shown in Fig. 1(a) change with the frequencies as shown in Fig. 1(b). The deepest contour in Fig. 1(b) represents the largest magnitude.

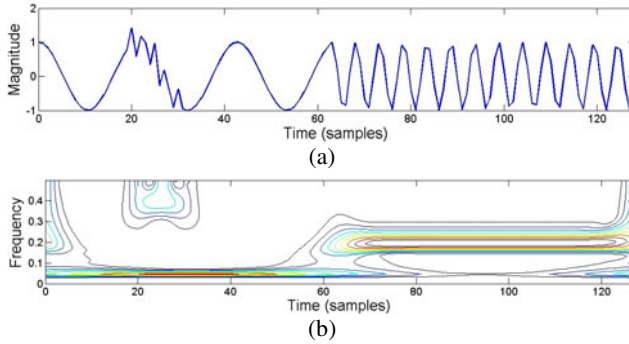


Fig. 1. (a) A chirp function with a high, medium and low frequencies, (b) contour plot of S transform of (a)

In nut shell, a major advantage of the Stockwell transform is that the modulating sinusoids are fixed with respect to the time axis. Such information lacks in other Transform methods viz. Fourier and Hilbert Transforms and also the sudden bursts in the signal or sudden change in the frequency is not best represented. However, the technique lacks frequency resolution in for typical input signals which can be clarified from the results presented in Section 4.

3 Hartley S-Transform

The Hartley transform (HT), like Fourier transform (FT) is an integral transform and also closely related to it. However, the difference lies in HT from the FT in that the forward and inverse transformations are identical and in that the HT involves only real operations [9]. As the decades have passed researchers have completely came up with many renewed HT algorithms and applications keeping in view of the advantages of HT. However, the very basic version of HT is given by eqn. (9) with $H(f)$ as the Hartley transform of the input signal $h(t)$.

$$H(f) = \int_{-\infty}^{\infty} h(t) * cas(2\pi ft) dt \tag{9}$$

where, $cas(2\pi ft)$ is given as eqn. (10)

$$cas(2\pi ft) = \cos(2\pi ft) + \sin(2\pi ft) \tag{10}$$

HT as given in eqn. (9) is used to find all the frequency components present in a given signal. However, it is not suited to find the local behavior of signals which have time-dependent spectral content and in this aspect it is similar to FT. In order to make HT suitable for a wide range of applications a short time Hartley transform (STHT) is also defined and is as shown in eqn. (11).

$$STHT(\tau, f) = \int_{-\infty}^{\infty} h(t)*w(t-\tau)*cas(2\pi ft)dt \tag{11}$$

However, it was observed that the STHT does provide time resolution of the spectral content, but fails to accurately resolve the low-frequency signals components of a signal when different ranges of frequencies are present. The reason being the window remains either narrow or broad which is a consequence of absence of dilation. As the Hartley and Fourier kernels have the same wavelength at the same value of f and thus STFT also suffers from the same problems as the STHT.

$$STFT(\tau, f) = \int_{-\infty}^{\infty} h(t)*w(t-\tau)*exp(-i2\pi ft)dt \tag{12}$$

Hence, for further improvement of both time and frequency resolution in STFT given as eqn. (12) the method described in Section 2, eqn. (4) has been proposed by Stockwell. A clear observation of eqn. (4) gives us an idea that ST is just STFT with the factor $exp(-i2\pi ft)$ remaining same but with the dilation of window with frequency f . And thus the ST may be rewritten as eqn. (13).

$$S(\tau, f) = \int_{-\infty}^{\infty} h(t)*w(t-\tau, f)*exp(-i2\pi ft)dt \tag{13}$$

The idea of HST originates from this point of discussion that by replacing the phase correction factor $exp(-i2\pi ft)$ with $cas(2\pi ft)$ produces the HST and is as shown in eqn. (14).

$$SHA(\tau, f) = \int_{-\infty}^{\infty} h(t)*w(\tau-t, f)*cas(2\pi ft)dt \tag{14}$$

The SHA matrix obtained in eqn. (14) is used to find the HST of a given input signal $h(t)$. The HST derived in eqn. (14) is used for analysis for the same input signal shown in Fig. 2(a). The improvement in frequency resolution by can be proofread from Fig. 2(b). High frequency is exactly represented compared to Fig. 1(b) with reduction in leakage. The low frequency strip extends from 0 to 128th sample in Fig. 1(b) though the low frequency is present only upto 64th sample in the input as shown in Fig. 1(a). Whereas, using HST the low frequency is represented upto 80th sample thus showing an improvement. The medium frequency components are also represented with best resolution without any cross terms. Thus HST is a better transformation technique in terms of frequency resolution compared to ST.

The order of the matrix obtained using eqn. (14) is same as that of the ST matrix in eqn. (4), i.e. no. frequency samples X no. of time samples. The matrix thus obtained is used to analyze various power system transient disturbances and is also compared with ST technique. Various case studies considered and results obtained have been presented in the next section.

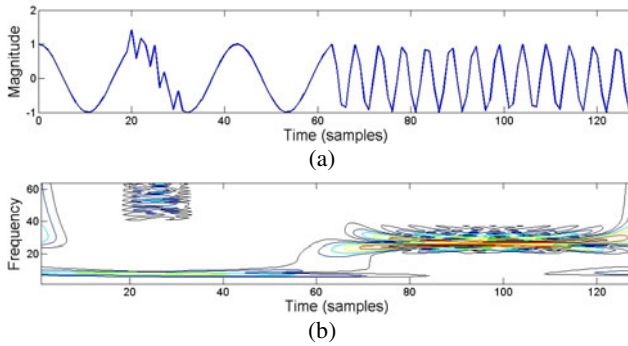


Fig. 2. (a) A chirp function with a high, medium and low frequencies, (b) contour plot of HS transform of (a)

4 Case Studies and Discussions

Some typical cases which are the sources of transients are considered to show the effectiveness of the technique used. These are very practical in nature and are of huge importance, as such transients causes stresses on the power system components. The protective devices are to be designed depending on the nature of the disturbance. In the absence of practical data the power system in the present work has been modeled in MATLAB/SIMULINK [12] and the transient data thus obtained has been utilized for determining the S and HS-contours. All the voltage magnitudes have been taken as 1p.u. with 50Hz frequency and the time interval being 0.2s. The training/input data has been sampled at a frequency of 5kHz thus giving 1000 samples. Different cases considered are as presented below.

4.1 Capacitor Switching

Power system deals with sudden switching of capacitors for various applications and thus produces transients. The transient nature arises due to the property of capacitor that it restricts itself to react for instantaneous change in voltage. The analysis of such a case is of great concern and here it has been considered as a case study. A voltage deviation is seen due to such transients. Hence, voltage waveform with a deviation of 5% in the load side has been taken as input for analysis. The deviations have been achieved by changing the capacitance and load values. The circuit breaker operation is carried out for 4 cycles with initially open condition for the voltage deviation case [7]. This type of behavior is practically seen in power system and is also of very huge importance as it creates stresses on the equipments. Its significance also extends for the design of the protective elements.

Transient disturbance created from the simulation circuit having 5% deviation in the voltage from $t=0.02s$ and having a very high frequency variation is shown in Fig. 3(a). This instant is very well captured in the S-contours separately at the high frequency range and rest of the signal which is a low frequency signal shows no such variation in Fig. 3(b). The bottom most contour strip belongs to fundamental frequency which is 50Hz in real time and 0.01Hz in ST domain. The deepest contour

with least area in the contour group in the top most area represents the high value of the frequency change nearly from 0.1s. It is seen here that the contours in the upper half frequency is extending upto 575th sample. As the sampling frequency is 5kHz the real time equivalent of this sample is 0.115s in the input waveform. However, the high frequency transient is ending before 0.115s in the input waveform and thus the ST contour plot is showing a leakage in frequency resolution.

In the case of HST contour plot shown in Fig. 3(c) it is seen that the contours in the upper half frequency are extending only upto 550th sample which is equivalent to 0.110s in real time. The signature obtained leads to a visual satisfaction that the identification is accurate. There is no unnecessary leakage. At the same time the contours are exactly ending up at 400Hz in the HST domain in y-axis which shows that the frequency is definite and maximum at 400Hz in HST domain. This type of accurate information and identification of transient events is of vast importance.

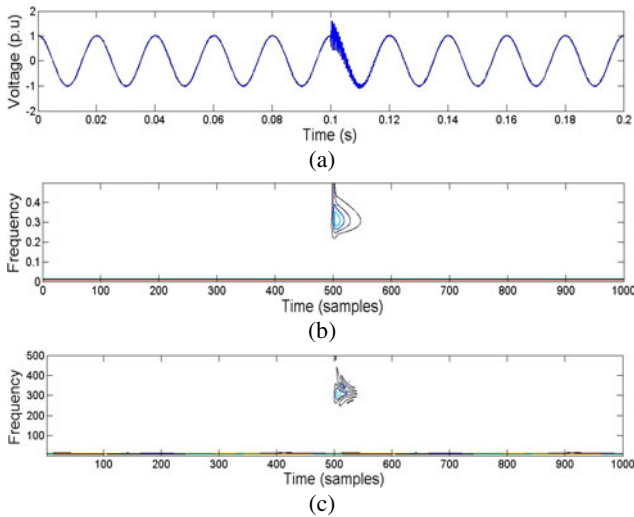


Fig. 3. (a) Voltage waveform obtained due to capacitor switching, (b) contour plot of ST of (a), (c) contour plot of HST of (a)

4.2 Bank to Bank Switching of Capacitors

Previously, a single capacitor which is causing a transient due to sudden switching has been considered as a case study. Here, a transient obtained due to bank to bank switching of capacitors is considered. It is well known fact that the capacitors are installed for sustainable operation of power systems. However, due to dynamic nature of the power system there is a huge variation of load in a day. Hence, in order to meet such variations the power system has to be supplied a controlled generation and will cause voltage instability. As capacitor banks have voltage level maintenance as one of the application they are switched bank to bank to provide more adaptive voltage support [1]. In this process the voltage feels a transient variation for an erstwhile

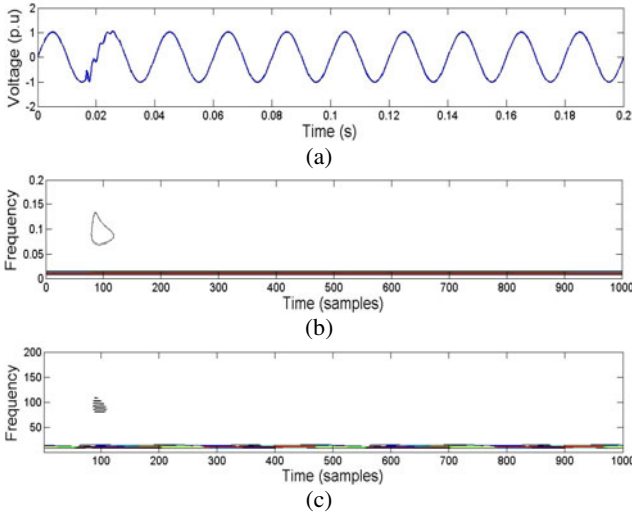


Fig. 4. (a) Voltage waveform obtained due to bank to bank switching of capacitors, (b) contour plot of ST of (a), (c) contour plot of HST of (a)

which will be in limits as specified. However, a maloperation may take place due to some irregularity and such an event has to be considered for analysis.

Fig. 4(a) shows the waveform obtained from an equivalent simulated model of a bank to bank capacitor switching. The switching is done by controlling the circuit breaker operation from $t=0.018s$ to $t=0.028s$. It is seen that a small deviation has occurred in the voltage exactly at the same time of operation of the circuit breaker. The nature of the disturbance is not known from the basic monitoring view. However, the analysis using the proposed technique can be done in terms of frequency ranges.

Fig. 4(b) shows the contour plot obtained using the proposed ST technique. It is seen that for the fundamental frequency there is a continuous strip at the bottom most of the plot with the same value of 0.01Hz in ST domain. However, there is a small area plotted with contours at the frequency range of 0.05-0.15Hz in the ST domain. Only one contour line is seen in this region. However, there are more oscillations in the input waveform which are not visualized exactly. Now let us move on to the analysis of the signal with the proposed HST technique.

Fig. 4(c) shows the HST contour plot of the waveform of Fig. 4(a). The behavior of the contour plot is an exact replica in time-frequency domain. The transient produced is of medium range and is represented exactly within 75Hz to 125Hz in HST domain. The contours are confined only to a specific time samples thus showing the quick damping of the transient oscillation. The contours are completely filled with blue thus clearly specifying the region of oscillations. Hence, the proposed HST technique is good in both time and frequency resolution. The analysis is visually good and its properties will be clearer from the further discussions on various case studies.

4.3 Voltage Spike due to Capacitor Switching

A transient event is basically classified as oscillatory transient and impulsive transient. The cases discussed in all the previous cases fall under oscillatory transients category and by its property the frequency of oscillations are high for a given time interval. In impulsive transient the magnitude can be either positive or negative [1]. The magnitude is of very high range and sustains only for a very short interval compared to oscillatory transient. A voltage spike is one of the examples of impulsive transient. The capacitor switching is one of the reasons. Unlike the previous cases the frequency components will be present only for 1-3 cycles. A waveform similar to voltage spike obtained from the simulation due to sudden ON/OFF operation of a capacitor is shown in Fig. 5(a). The spikes are there both in positive and negative polarity. Its analysis and identification using the ST and proposed HST technique is shown in Fig. 5(b) and 5(c) respectively.

It is seen that the contours are seen from a low frequency to high frequency i.e. from 0.05 to 0.4Hz and 50-500Hz in ST and HST domain respectively. Such signature gives an idea that the range of frequencies are varying exactly at that time instant. The contours from 475th sample to 500th sample are continuous in Fig. 5(b) while those are discontinuous in Fig. 5(c) for the same time samples. The difference in the representation carries useful information. The discontinuity depicts that all the frequencies present have various magnitude whereas such information is lacking in ST technique. The nature magnitudes of various frequencies are of paramount importance to the monitoring point so as to take the precautionary actions. However, the second contour set from 500th sample to 550th sample in Fig. 5(b) is continuous and having same color throughout. The prediction is that the frequency components present have almost same magnitude which is also same in Fig. 5(c). It is seen that the contours are on the positive direction for both negative and positive polarity spikes.

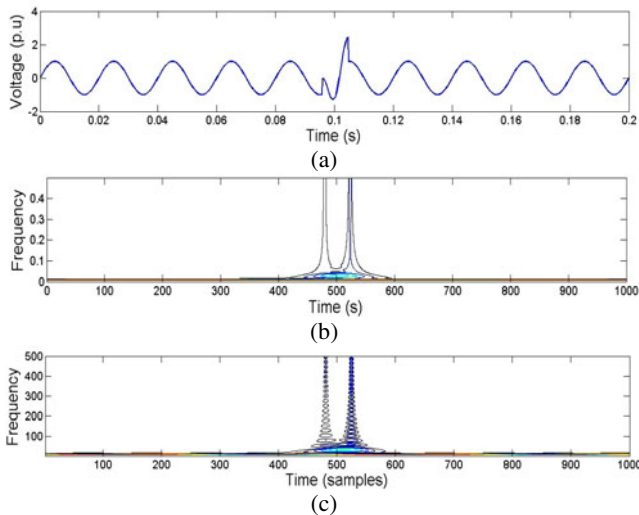


Fig. 5. (a) Voltage spike obtained due to switching of capacitor, (b) contour plot of ST of (a), (c) contour plot of HST of (a)

4.4 Switching of Wind Plant

As the wind may not flow continuously and there may be a sudden change in the wind speed a transient arises due to switching of a wind turbine with the power grid. Such a typical case has been considered and analyzed here. The voltage measured at the point of common coupling (PCC) is shown in Fig. 6(a). The voltage and frequency during transient are high compared to normal operating conditions. Using the ST contours of the input voltage is obtained and the plot is as shown in Fig. 6(b). From these contours a sudden change in frequency due to transient when the wind generator suddenly turned ON can be clearly visualized. Exactly at time sample 400 there are contours with spike structure from low to high frequency. There is a side lobe within the frequency range 0.1 to 0.2Hz and between 400-470th sample in ST domain. This lobe corresponds to oscillations in input from 0.084 to 0.1s. However, the oscillations are not correct represented with the ST technique. The uniqueness and accuracy of HST technique can be visualized from HST plot in Fig. 6(c).

The contour representation is nearly same as CST plot in 6(b). However, the lobe is exactly stretching upto 490th sample which is showing that the oscillations are prolonged upto 0.098s. The information obtained is very accurate compared to ST. It can be seen that the red color in the HST plot is only at the highest frequency which shows that the magnitude of high frequency is large and the representation is accurate. Such information is very useful to the utilities for taking preventive actions. The signature obtained is very unique. The uniqueness is not only maintained in this case but can also be verified from all the HST contour plots obtained till now. The efficacy of the technique is tested on other case studies and due limit in number of pages those case studies are not reported at present.

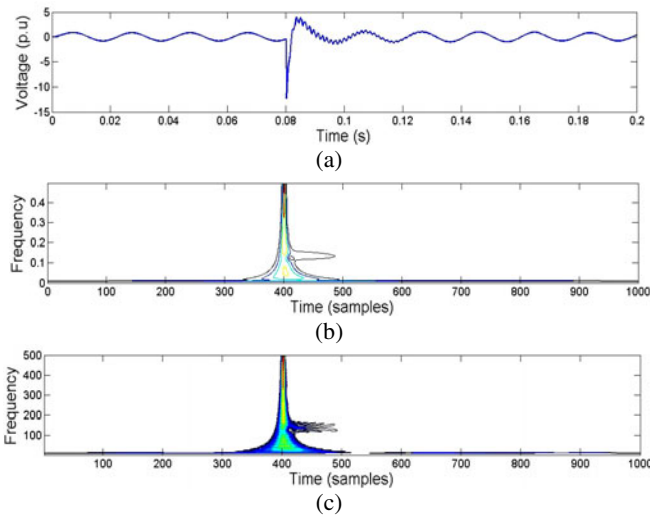


Fig. 6. (a) Voltage waveform obtained due to switching of wind plant into a power grid, (b) contour plot of S transform of (a), (c) contour plot of HS transform of (a)

5 Conclusions

The modified Stockwell transform termed as Hartley S-transform is used in this paper for detection of sudden transients in power system. The time–frequency plot of the HS–transform has a significant potential in comparison to S-transform. It is noted that the frequency dependant resolution of the HS–transform allows the detection of high-frequency bursts and shows good frequency resolution on the long period signal. The present work shows that HST provides an interesting and significant tool in detecting transient problems in power systems and is better than ST.

References

1. Greenwood, A.: *Electrical Transients in Power Systems*, 2nd edn. John Wiley & Sons, USA (1923)
2. Lobos, T., Rezmer, J., Koglin, H.J.: Analysis of power system transients using wavelets and Prony method. In: *IEEE Porto Power Tech. Conf.* (2001)
3. Janıcek, F., Mucha, M., Ostrozlık, M.: A new protection relay based on fault transient analysis using wavelet transform. *Journal of Electrical Engineering* 58(5), 271–278 (2007)
4. Andrade, M., Messina, A.R.: Application of Hilbert techniques to the study of subsynchronous oscillations. In: *IPST International Conference on Power Systems Transient*, pp. 172–178 (2005)
5. Mamis, M.S., Abbasov, T., Herdem, S., Koksall, M.: Transient analysis of electrical machines by differential Taylor transform. In: *IPST International Conference on Power System Transient*, pp. 325–328 (1999)
6. Mo, F., Kinsner, W.: Wavelet modeling of transients in power systems. In: *IEEE Conf. Communications, Power and Computing*, pp. 132–137 (1997)
7. Dafis, C.J., Nwankpa, C.O., Petropuh, A.: Analysis of Power System Transient Disturbances Using an ESPRIT-based Method. In: *IEEE Conf.*, pp. 437–442 (2000)
8. Girgis, A.A., McManis, R.B.: Frequency domain techniques for modeling distribution or transmission networks using capacitor switching induced transients. *IEEE Power Engineering Review*, 74–79 (1989)
9. Pinnegar, C.R., Mansinha, L.: Time–frequency localization with the Hartley S-transform. *Journal of Signal Processing* 84, 2437–2442 (2004)
10. Ruirui, Z.S.L., Jeffers, W.Q., Heptol, T., Guimin, Y.: The research of power quality analysis based on improved S-transform. In: *The Ninth International Conference on Electronic Measurement & Instruments, ICEMI*, pp. 477–481 (2009)
11. Stockwell, R.G.: S-transform analysis of gravity wave activity from a small scale network, Ph.D. dissertation, Dept. of Physics, Western Ontario Univ., London, Ontario (1999)
12. MATLAB/ SIMULINK 7.6 version

A Modified Kolmogorov-Smirnov Correlation Based Filter Algorithm for Feature Selection

Pakkurthi Srinivasu¹, P.S. Avadhani², Suresh Chandra Satapathy³,
and Tummala Pradeep⁴

¹ Department of CSE, ANITS, Visakhapatnam

² Department of CS&SE, Andhra University, Visakhapatnam

³ Department of CSE, ANITS, Visakhapatnam

⁴ Department of CSE, BIT, Ranchi, Jharkhand

Abstract. A feature selection is a technique of selecting a subset of relevant features from which the classification model can be constructed for a particular task. Feature selection is a preprocessing step of machine learning which is effective in reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving results. In this paper, a modified Kolmogorov-Smirnov Correlation Based Filter algorithm for Feature Selection is proposed based on Kolmogorov-Smirnov statistic which uses class label information while comparing feature pairs. Results obtained from this algorithm are compared with two other algorithms, Correlation Feature Selection algorithm (CFS) and simple Kolmogorov Smirnov-Correlation Based Filter (KS-CBF), capable of removing irrelevancy and redundancy. The classification accuracy is achieved with the reduced feature set using the proposed approach with two of the standard classifiers such as the Decision-Tree classifier and the K-NN classifier.

1 Introduction

Feature selection is a preprocessing step of machine learning which is effective in reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving results. In recent years, data has become increasingly larger in both number of instances and number of features in many applications such as genome projects, text categorization, image retrieval, and customer relationship management. The increase of data and features cause serious problems to many machine learning algorithms with respect to scalability and learning performance. Hence, feature selection is very much important for machine learning tasks which include high dimensional data.

Feature selection evaluation methods fall into two broad categories, Filter model and Wrapper model [2]. The Filter model depends on characteristics of the training data to select some features without involving any learning algorithm. The wrapper model needs one predetermined learning algorithm in feature selection and uses its performance to evaluate and determine which features are to be selected. As for each new subset of features, the wrapper model needs to learn a hypothesis/ classifier. It tends to find features better suited to the predetermined learning algorithm resulting in superior learning performance, but it also tends to be more computationally expensive

and less general than the Filter model. When there are more number of features, the Filter model can be used because of its computational efficiency. Filters have the advantage of fast execution and generality to a large family of classifiers than wrappers [13].

Figure 1 provides a depiction of a simple classification process where a Feature Selection process that uses a filter is involved. The training and testing datasets after the dimensionality reduction process is fed to the ML (Machine Learning) algorithm. In this paper, we have employed a Filter model for the evaluation of selected features.

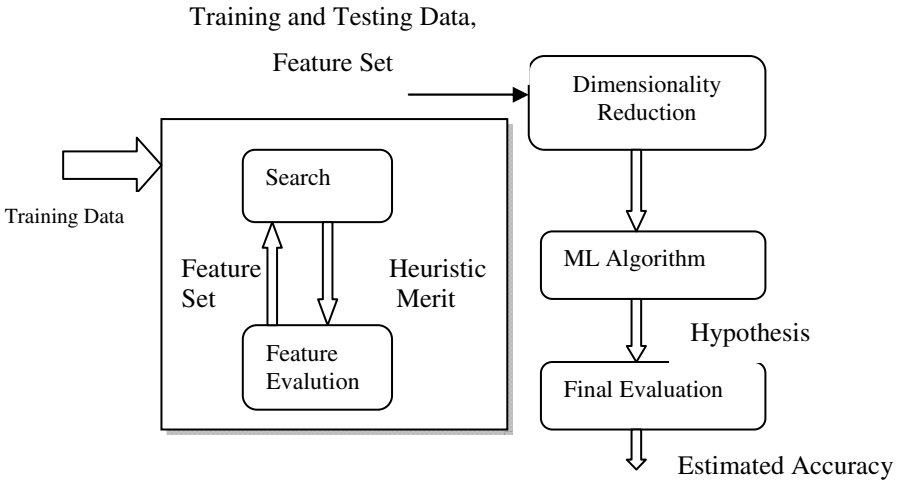


Fig. 1. Classification Process that involves Feature Selection stage with Filter approach

2 Theoretical Framework

2.1 Correlation Based Feature Selection

This algorithm [1] is based on information theory and uses symmetrical uncertainty (SU) as the filter for the evaluation of the selected feature set. This algorithm involves certain concepts such as mutual information [3], entropy, information gain and symmetrical uncertainty. The process used here in finding correlations between various attributes is different from that used in FCBF [11]. In the first step, it processes the given training dataset and initial feature set and removes all the irrelevant features by finding the strength of prediction of feature-to-class. In the second step, it uses this relevant feature set and training dataset to remove all the redundant features and finally presents the significant feature set that is well supervised and uncorrelated with other features.

Entropy: Entropy as given by Shannon is a measure of the amount of uncertainty about a source of messages [5]. The entropy can be described by when the variable Y before and after observing values of another variable X:

$$H(Y) = - \sum p(y_i) \log(p(y_i))$$

and

$$H(Y/X) = - \sum p(x_j) \sum p(y_i/x_j) \log(p(y_i/x_j))$$

Here $p(y_i)$ is the prior probabilities for all values of random variable Y and $p(y_i/x_j)$ is the conditional probability of y_i given x_j . Without any uncertainty if all features belong to the same class when the entropy is 0 by observing Y as classes and X as features in a data set. On the other hand, members in a feature set are totally random to a class if the value of entropy is 1. The range of entropy is between 0 and 1.

Information Gain

The amount by which the entropy of X decreases reflects additional information about Y provided by X and is called information gain, given by

$$\begin{aligned} \text{Gain, } I(Y; X) &= H(Y) - H(Y|X) \\ &= H(X) - H(X|Y) \\ &= H(Y) + H(X) - H(X, Y). \end{aligned}$$

However, information gain is biased if feature with more values[4], which is the features with greater numbers of values will gain more information than those with fewer values even if the former ones are actually less informative than the latter ones.

Symmetrical Uncertainty

Because of the limitation provided by the usage of Information gain, we use another heuristic called Symmetrical Uncertainty and is given by:

$$SU(Y; X) = 2[I(Y; X) / (H(X) + H(Y))]$$

The average of two uncertainty variables can be computed with symmetrical uncertainty and it compensates for information gains bias toward features with more values. It can be normalized its values in the range [0,1]. The digit 1 indicates that knowing the value of either one completely predicts the value of the other. A digit 0 indicates that X and Y are independent of each other.

Correlation Feature Selection Algorithm described as follows[1]:

1. Remove irrelevant features from the data set features
2. Input original data set D that includes features X and target class Y
3. For each of the feature X_i
Calculate mutual information $SU(Y; X_i)$
4. Sort $SU(Y; X_i)$ in descending order
5. Put X_j whose $SU(Y; X_i) > 0$ into relevant feature set R_{xy}
6. Remove redundant features from data set features
7. Input relevant features set R_{xy}
8. For each feature X_j
Calculate pair wise mutual information $SU(X_j; X_k)$ for all $j \neq k$
9. $S_{xx} = \sum (SU(X_j; X_k))$

10. Calculate means μ_R and μ_S of R_{xy} and S_{xx} , respectively

$$W = \mu_S / \mu_R$$

11. $R = W \cdot R_{xy} - S_{xx}$

12. Select X_j whose $R > 0$ into final set F

2.2 Kolmogorov-Smirnov Test

Equivalence of two random variables may be evaluated using the Kolmogorov-Smirnov (KS) test [12]. In statistics, the Kolmogorov-Smirnov test (K-S test) is a nonparametric test for the equality of continuous, one-dimensional probability distributions that can be used to compare two samples (two-sample K-S test).

The **empirical distribution function** F_n for n independent and identical observations X_i is defined as:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I_{X_i \leq x}$$

where $I_{X_i \leq x}$ is the indicator function, equal to 1 if $X_i \leq x$ and equal to 0 otherwise.

The following are the steps using KS test: (KS test for two features F_i, F_j)

1. Discretization of both features F_i, F_j into k bins.
2. Estimate the probabilities in each bin.
 - 2.1. Calculate the cumulative probability distributions for both features F_i, F_j .
 - 2.2. Calculate KS statistic

A two-step Kolmogorov-Smirnov Correlation Based Filter (K-S CBF) algorithm. [7][8][9][10]

Relevance analysis

1. Calculate the $SU(X,C)$ relevance indices and create an ordered list S of features
2. according to the decreasing value of their relevance.

Redundancy analysis

3. Take the feature X from the S list
4. Find and remove all features for which X is approximately equivalent according to the K-S test
5. Set the next remaining feature in the list as X and repeat step 3 for all features that follow it in the S list.

3 Modified KS-CBF Test

The proposed algorithm introduces the concept of binning the input dataset into n bins. Redundancy is calculated in each of the bins individually and the set of redundant features for each bin are stored. Finally, the set of features that are common in all the bins (here, it can also be taken as an input parameter to decide at run-time as required) are considered as redundant for the input dataset and they can be eliminated.

In each bin, again we divide the available set of records into a particular number of partitions for which the actual KS-test is applied. The set of redundant features in each of these partitions is mixed up with those for the other partitions in that bin. So, now we can expect a good redundant subset to be produced from each bin. The union operation ensures that the possibility of redundancy has been checked for every feature in its entirety. The intersection operation ensures that a feature which is actually non-redundant will not be claimed as being redundant. Now, since we perform an intersection of the redundant features obtained from all the bins, the final feature subset produced will not contain these features and could sufficiently represent a significant subset of features that can be used for the classification process.

Modified K-S CBF Algorithm

Relevance analysis

1. Order features based on decreasing value of SUC (f, C) index which reflects the decreasing value of their relevance.

Redundancy analysis

2. Pass the dataset with relevant features for KS test measure.

3. Discretize the dataset into n bins each containing approximately same number of records.

4. For each of the bin B_i

4.1. Form k data partitions each approximately containing the same number of records.

4.2. For each of the k partitions P_1, P_2, \dots, P_k

4.2.1. Initialize F_i with the first feature in the F-list.

4.2.2. Find all features for which F_i forms an approximate redundant cover using K-S test.

4.2.3. Set the next remaining feature in the list as F_i and repeat above step for all features that follow it in the F list.

4.3. Take the union of all these redundant features into B_i

5. Get the common features that are redundant in all bins.

6. Remove those features and get the significant subset of features.

A Modified Kolmogorov-Smirnov Correlation Based Filter Algorithm

In the first step, a Symmetrical Uncertainty filter has been applied to remove the irrelevant features. The dataset containing this filtered subset is now passed to the second step to perform the redundancy analysis and obtain the set of redundant features. These redundant features are removed from the remaining feature set and finally the dataset with significant features is only considered for further analysis.

In most cases, classification accuracy using this reduced feature set produced equaled or bettered accuracy using the complete feature set. However, the dimensionality of the feature set has been reduced to a better extent than compared to the simple KS-CBF algorithm. This gives a good computational gain over the simple KS-CBF when testing with a classifier as the new feature set contains less number of dimensions. In few cases, it is however observed that the proposed algorithm has been producing results that are quite less efficient when tested with a classifier than that

compared with the simple KS-CBF results. However, this can still be handled and improved by varying the values of number of bins which is taken as an input parameter. Feature selection can sometimes degrade machine learning performance in cases where some features were eliminated which were highly predictive of very small areas of the instance space. The size of the dataset also plays a role in this and as we are further dividing into smaller bins, it sometimes affects the process when there are no sufficient records in a bin to perform the test. Also, when the information available in the dataset is very random with a large range of distinct feature values, then also this algorithm could produce very good results than others.

It has been observed that, as the number of bins during the test is increased, the number of redundant features is increasing and the final feature set produced is getting smaller. However, in some cases, this has been leading to a slight decrease in detection rates. Yet, the performance gain obtained is very much high. So, the effect of slight decrease in detection rates can be negotiated with the rapid increase in performance.

4 Experimental Setup and Results

Table1 illustrates the datasets worked upon along with their peculiar properties. Tables2 and Table3 give the set of features selected with various feature selection algorithms. Table4 and Table5 give the detection rates with two major classification algorithms Decision-Tree and K-NN algorithm.

Table 1. List of datasets used in this experiment

Dataset	No. of Features	No. of Instances	No of Classes	Class Distribution
Ionosphere	34	351	2	16/25
Pima	8	768	2	500/68
Wdbc	31	569	2	212/357
Wine	13	178	2	59/71
Dos-Normal	41	4765	2	2371/2394
Probe-Normal	41	4346	2	1996/2350
U2R-Normal	41	326	2	52/274
R2L- Normal	41	2137	2	1062/1075

Table 2. Selected Features of KDD 99 datasets

Data Set	No. of features	Correlation Feature Selection	Simple KS-test	Modified KS-test
Normal-DoS	41	1,5,10,23,24,25,, 31,33,35,38,39	2 - 6, 12, 23, 24, 31- 37,	2 - 6, 12, 23, 32, 36
Normal-Probe	41	4,24,27-32,40,41	3-6, 12, 23, 24, 27, 32-37, 40	3 - 6, 12, 23, 32, 33, 37
Normal-U2R	41	1,10,13,14,17,23, 35-37	1,3,5,6,32-34,24,36,37	1, 3, 5, 6, 24, 32 - 34, 36
Normal-R2L	41	1,10,22,23,31,32, 34-37	1,3,5,6,10,22-24,31-37	2-6, 24, 32, 33, 36, 37

Table 3. Selected Features sets for various UCI datasets

Dataset	No. of features	Correlation Feature Selection	Simple KS-test	Modified KS-test
Ionosphere	34	6, 12, 22, 24, 25, 27, 29, 32 - 34	4, 25, 28	4, 25, 28
Pima	8	2, 5, 6, 8	2 - 8	2,6,7
Wdbc	31	8, 9, 11, 18, 21, 27 - 31,	1, 4, 5, 8, 9, 12, 14 -18, 21, 24, 25, 28	1, 4, 5, 8, 9, 14 -16, 18, 19, 21 - 23, 28
Wine	14	1, 2, 6, 9, 10, 12	1, 2, 7, 10, 12, 13	1, 2, 7, 10, 12, 13

Table 4. Detection Rates with various feature subsets for KDD99 Datasets

Data Set	K-NN Classifier				Decision Tree			
	Full Set	Correlation Feature Selection	Simple KS-test	Modified KS-test	Full Set	Correlation Feature Selection	Simple KS -test	Modified KS-test
Normal - DoS	99.92	99.55	99.73	99.39	99.40	99.73	99.41	99.41
Normal - Probe	97.37	91.5	97.28	97.28	100	90.44	100	100
Normal -U2R	96.80	92.0	93.89	99.20	100	95.2	100	100
Normal -R2L	99.28	83.73	98.40	98.92	97.72	94.2	97.37	98.95

Table 5. Detection Rates with various feature subsets for UCI Datasets

Data Set	K-NN Classifier				Decision Tree			
	Full Set	Correlation Feature Selection	Simple KS-test	Modified KS-test	Full Set	Correlation Feature Selection	Simple KS-test	Modified KS-test
Ionosphere	80.82	87.67	84.24	84.24	86.98	84.24	69.18	69.36
Pima	74.02	75.32	71.42	75.65	59.7	54.5	50.0	56.49
Wdbc	67.36	92.05	67.36	67.78	85.77	85.35	93.31	93.61
Wine	95.55	99.77	96.65	97.77	99.33	100	97.78	97.78

It can be seen that the proposed approach selects a less number of significant feature subset in many cases than the other two algorithms. Also, the accuracy and efficiency of this approach was much better in most of the cases. In few cases, the accuracy slightly reduced but it is not that much far away from the other methods and this method outperformed both the CFS and KS-test in terms of efficiency i.e. in terms of execution performance. As compared to the results in [4] and [6], the results obtained in our experiment are good and encouraging.

5 Conclusion

A modified Kolmogorov-Smirnov Correlation Based Filter algorithm for Feature Selection is proposed in this paper. The proposed algorithm has the computational demands that are very much similar to the traditional KS-test and is proportional to the total number of bins. A comparative test with some widely used feature selection algorithms showed its better performance in terms accuracy. The statistical factor of 0.05 has been used in test which can be varied. The number of bins and number of partitions should be given depending on the number of records in the incoming dataset. According to our observations, good results are obtained if a bin is made to contain at least 40 records.

Since a filter approach is used, the results can be well suited to any classifier process. The results in our experiment have been tested with Decision Tree classifier and K-NN classifier. Various variants of the Kolmogorov-Smirnov test exists and the algorithm may be used with other indices for relevance indication.

References

1. Chou, T., Yen, K., Luo, J., Pissinou, N., Makki, K.: Correlation Based Feature Selection for Intrusion Detection Design. In: IEEE Military Communications Conference, MILCOM 2007, pp. 1–7 (2007)
2. Hall, M.A., Smith, L.A.: Feature subset selection: A correlation based filter approach. In: Proc. Intl. Conf. Neural Inform. Processing Intell. Inform. Syst., pp. 855–858 (1997)
3. Bonev, B., Escolano, F., Cazorla, M.A.: A Novel Information Theory Method for Filter Feature Selection. In: Gelbukh, A., Kuri Morales, Á.F. (eds.) MICAI 2007. LNCS (LNAI), vol. 4827, pp. 431–440. Springer, Heidelberg (2007)
4. Chou, T.-S.: Ensemble Fuzzy Belief Intrusion Detection Design Thesis. Florida International University, Miami (2007)
5. Bancarz, I.: Conditional Entropy Metrics for Feature Selection, University of Edinburgh, College of Science and Engineering, School of Informatics (June 2005)
6. Blachnik, M., Duch, W., Kachel, A., Biesiada, J.: Feature Selection for Supervised Classification: A Kolmogorov-Smirnov Class Correlation-Based Filter. In: AIMeth, Symposium On Methods Of Artificial Intelligence, Gliwice, Poland, (November 10-19, 2009)
7. Duch, W., Biesiada, J.: Feature Selection for High-Dimensional Data: A KolmogorovSmirnov Correlation-Based Filter Solution. In: Advances in Soft Computing, pp. 95–104. Springer, Heidelberg (2005)
8. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 2nd edn. (2006)
9. The Kolmogorov-Smirnov Test When Parameters are estimated from data: Hovhannes Keutelian, Fermilab
10. Webera, M.D., Leemisa, L.M., Kincaida, R.K.: Minimum Kolmogorov-Smirnov test Statistic Parameter Estimates. Journal of Statistical Computation and Simulation 76(3), 195–206 (2006)
11. Yu, L., Liu, H.: Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. In: Proceedings of the Twentieth International Conference on Machine Learning, Washington, D.C, pp. 856–863
12. Biesiada, J., Duch, W.: Feature Selection for High-Dimensional Data: A Kolmogorov-Smirnov Correlation-Based Filter. In: CORES, pp. 95–103 (2005)
13. Hall, M.A.: Correlation-based Feature Subset Selection for Machine Learning. PhD dissertation, Department of Computer Science. University of Waikatoa (1999)

Software Test Effort Estimation Using Particle Swarm Optimization

Prasanta Bhattacharya¹, Praveen Ranjan Srivastava¹, and Bhanu Prasad²

¹ Department of Computer and Information System
Birla Institute of Technology and Science,
Pilani 333031, Rajasthan, India

² Department of Computer and Information Sciences,
Florida A&M University, Tallahassee, Florida 32307, USA
{h2009120,praveenr}@bits-pilani.ac.in,
prasad@cis.famu.edu

Abstract. Software testing is a key component in the software development life cycle. This paper presents a modification to the Constructive Cost Model (COCOMO) technique by using particle swarm optimization. The resultant technique significantly increases the accuracy of the COCOMO approach and also incorporates the much needed flexibility related to the software and the development team.

Keywords: Software testing, software testing effort (STE), particle swarm optimization (PSO), COCOMO, test effort drivers (TED).

1 Introduction

Software engineering is a dedicated study and a systematic approach for producing software of better quality, low cost and higher efficiency which can additionally be built faster and maintained easily [1]. Software testing is an important component in the software development life cycle and aims at finding errors and defects in the developed software [2, 3]. During the testing process, validation and verification of the software is conducted to check whether the software meets the requirements that guided its analysis, design and development. It is fundamentally true that efficient testing leads to good quality software, user satisfaction, and lower maintenance cost. It is due to this implied criticality that nearly 35% of the elapsed time and more than 50% of the total software development cost are expended on the testing process [3, 4, 5, 6, 12]. The process of estimating the Software Development Effort (SDE) is completely based on uncertain and/or noisy inputs [4, 7] and this is the reason why most software projects today tend to face the effort estimation problem. The SDE can be defined as the effort required to develop and maintain software and is often the result of many effort estimation [8] and prediction models, including the one described in this paper. Software Testing Effort (STE) is generally around 40-50 percent of SDE [5, 9]. STE is generally estimated as a percentage of the calculated SDE, based on certain heuristics and previous experiences. STE cannot be determined independently of SDE estimation and there is no standard procedure presently to determine an accurate value for it.

The STE is calculated as a certain percentage of SDE, depending on the "criticality" of the project and the organizational practices [6, 12]. To calculate the required STE, we follow a number of steps: first, the overall SDE is estimated using any of the models discussed above and then it is "weighted" with a confidence factor, which is a heuristic to take care of the programmer or team capabilities. The value of this heuristic varies widely and is usually based on prior experiences with similar projects. Till date, there is no standard procedure to determine an exact and universally acceptable value of this heuristic. This research addresses the STE estimation problem by proposing a model that is a combination of the Constructive Cost Model (COCOMO), Particle Swarm Optimization (PSO), conventional weighing techniques, and the existing Test Effort Drivers (TEDs) (TEDs are explained in detail under Section 5). In this model, various TEDs have been appropriately weighted to reflect their need or criticality in the given project and the weights are optimized using the PSO technique for fast and accurate convergence. The proposed model also takes into account the confidence level (C) of the developer or tester to obtain an accurate estimation of STE.

2 Related Work

The design of software that is optimal in terms of time, cost, efforts and other resources is very important and necessary in software engineering. A systematic formulation of STE is needed to ensure that the above optimization constraints are fully satisfied before the release of the software. Estimating the cost and duration of STE is a major challenge these days. An early estimation of STE is based on the testing metrics, which generally overestimate the efforts, depending on the expertise of the software testing team [10]. Halstead has developed a set of metrics to measure the complexity of a program module directly from its source code [13, 14]. Kushwaha and Misra [15] have used a cognitive information complexity method to estimate the STE. Nageswaran [6] presented a method for estimating STE to perform all functional test activities based on the use case points. Jorgensen [17] has emphasized the importance of human factors in SDE estimation. An estimation model for test execution effort based on the test specifications was proposed by Aranha and Borba [16]. An approach for the development of SDE and schedule estimation models using soft-computing techniques was first presented by Sheta et al. [23]. Dawson [18] illustrated a neural network theory for STE estimation. Srivastava et al. [19][20], in their earlier work, have proposed multiple approaches for STE estimation by integrating Halstead matrices with fuzzy logic.

The current research builds upon the ongoing research on the optimization of STE estimation using soft computing and presents a novel model using the PSO algorithm. The rest of the paper is organized as follows. Section 3 provides an overview of COCOMO. Section 4 describes the PSO heuristic and the various parameters involved. Section 5 briefly explains various TEDs that were selected for the course of this research. Section 6 presents the algorithm for estimating STE and Section 7 discusses the analysis of the simulation results of the algorithm. Finally, Section 8 illustrates the future scope of this work.

3 An Overview of the Constructive Cost Model

Several cost estimation techniques for software development have been developed and they are broadly grouped into two major categories: algorithmic models and non-algorithmic models [8, 11]. Constructive Cost Model (COCOMO) [8, 11] is regarded as one of the best documented algorithmic cost-estimation models based on the number of lines of code written. The intermediate COCOMO model computes SDE as a function of the program size and a set of project cost drivers. Each of these attributes receives a rating on a six-point scale that ranges from "very low" to "extra high", in the increasing order of importance or value. The product of all the effort multipliers results in the effort adjustment factor (EAF), the typical values for which range from 0.9 to 1.4. The SDE can then be calculated by using the following formula [8, 11]:

$$\text{SDE} = a_i * (\text{KLOC})^{b_i} * \text{EAF} \quad (1)$$

Where, KLOC is the total lines of code in the software modules, a_i and b_i are constants which vary depending on the type of the implementing organization [8]. Clearly, there are two main disadvantages in calculating the SDE, and subsequently the STE, using the above COCOMO model. First, the model makes assumptions on the form of the effort calculating function that consists of some known constants which are dependent on varying organizational conditions. Second, the model is adjusted or modified according to certain local factors. These factors are evaluated using qualitative values such as 'very low', 'complex', 'important' and 'essential'. When dealing with such subjective assessments in calculating the attributes, imprecision and inaccuracy in results are unavoidable.

4 Overview of Particle Swarm Optimization Heuristic

Particle Swarm Optimization (PSO) is a stochastic optimization technique that borrowed its inspiration from the social behavior of birds flocking or fish schooling [21, 22]. Even though PSO shares many similarities with traditional evolutionary computation techniques such as Genetic Algorithms (GA), it outperforms most evolutionary techniques in terms of efficiency and resource utilization. The system is initialized with a population of random solutions and searches for global optima by updating the generations. However, unlike GA, PSO has no evolution operators such as crossover and mutation and does not require any kind of binary encoding. In PSO, the potential solutions called particles fly through the problem space by following the current optimum particles. Since, in PSO, the number of factors to be tuned are much lower than that of any other stochastic optimization methods, it has emerged as a first rate choice for many researchers. We commonly employ PSO based optimization techniques to the area of score fusion and attempt to find the optimal weights, w_1 and w_2 for which the score fusion (S) is maximized or minimized.

$$S = w_1 * \text{score}_1 + w_2 * \text{score}_2 \quad (2)$$

Here, score_1 and score_2 are Tier 1 and Tier 2 scores respectively and they would be obtained from the matching procedures. The PSO implementation provides us with the optimal weight set $[w_1, w_2]$ for which the score can be maximized. The main idea

behind the score fusion process is to generate the maximal scores for a genuine subject and to generate a minimal score for an incorrect one. For every iteration of the swarm simulation, the particle keeps track of three parameter values [21, 22] namely $present[]$ which denotes its present position in the neighborhood, $pbest[]$ which denotes the best value of the fitness function it has found so far and $gbest[]$ which denotes the best value found by any particle thus far. These three parameters are updated by using the following two equations [21, 22]:

$$v[] = inertia * v[] + c_1 * rand() * (pbest[] - present[]) + c_2 * rand() * (gbest[] - present[]) \quad (3)$$

$$present[] = present[] + v[] \quad (4)$$

where, $v[]$ denotes the particle velocity directed towards the solution in the solution space, c_1 and c_2 are learning factors assumed generally to be 2 and inertia factor which is assumed to be 1. The $rand()$ function generates a random number within a problem-dependent bounded space. In perspective of the software testing domain, the 'inertia' factor accounts for the resistance to the change in work practice or work nature of the programmer when we are using the PSO to evaluate the confidence level of the programmer. The learning factors provide a controlling weight to the confidence levels as well as the test effort drivers and keep these within permissible limits.

5 Test Effort Drivers

Various factors on which STE depends are commonly known as Test Effort Drivers (TEDs). In this research, three such TEDs have been identified as described in this section viz.

- (a). Software complexity (SC): If the complexity of the entire project is very high, the amount of STE should be increased as the number of test cases will also be high.
- (b). Software quality (SQ): The software quality can be measured using several factors like functionality, reliability, usability, efficiency, and transferability. If the software quality needs to be high, the values of these parameters must be kept high, which in turn will result in a higher STE.
- (c). Work force drivers (WFD): These include the tester capability, programmer capability, experience with application domain, programming language paradigm, exposure and experience with a given language, use of modern programming practices, use of software tools, degree of dependency on external tools and others.

6 Algorithm for STE Estimation

The algorithm for estimating the STE by integrating the existing COCOMO approach with PSO model is presented as follows:

Step 1. Prepare or gather functional and non-functional requirements for the project. Prepare a software requirement specification document for the project or module under consideration.

Step 2. Estimate the KLOC for the proposed software or project from the software requirement specification document.

Step 3. Define Confidence Level (C) for the lead developer or development team as a function of the key deciding attributes. More specifically, estimate the factors influencing the value of C [21] (Note: to get a SDE estimation, organizations generally depend on the project manager and this is the reason why some projects are either over budgeted or under budgeted, or have a problem at the time of their release. To address this issue, organizations calculate the value of C of the project manager to get a close estimate of SDE. The value of C is calculated on a [0, 1] scale, i.e., $0 \leq C \leq 1$). The value of C depends on the following factors:

- Weighted mean project-experience of the team (measured in number of years).
- Team capability (measured as the success rate of the project manager, normalized to a number between 0-10) along with a weighted mean of the project analyst capabilities and programmer capabilities.
- Familiarity of the project team (measured as the number of similar projects they have done in the past).
- Use of software tools and disciplined methods (rated on a scale of 0-10).

Note that these factors can be altered as per the needs of the testing team but the methodology remains the same.

Step 4. Optimize the weights for the key attributes using PSO and obtain an accurate value of C.

Step 5. Calculate the updated (i.e., refined) approximation of KLOC as follows: $KLOC (i.e., updated KLOC) = KLOC (obtained in Step 2) * C$.

Step 6. Calculate the SDE using the updated KLOC value by following the COCOMO approach.

Step 7. Identify the TEDs which impact the testing process, and express the relationship and criticality of TEDs using a relation to obtain the value of the Parentage (P) of STE in SDE.

Step 8. Optimize P using PSO.

Step 9. Calculate the approximate value of STE as follows: $STE = P * SDE$.

Step 10. Exit.

7 Analysis of Simulation Results

The algorithm presented in previous section was validated for a fictitious software project developed by a large scale organization and the results are presented in this section. As explained in the algorithm, there are two entities that are optimized using the PSO algorithm viz. the values of C and P. In our case, we assumed that the C value of the development team is expressed as a weighted function of the team experience (E), familiarity with project (F) and team morale (M). While the first two attributes account for the influence of historical factors on C, the last attribute accounts for the influence of more recent incidents on C. The assumed relation may be written as follows:

$$C = (E_i - a_i) * (F_i - b_i) * (M_i - c_i) \quad (5)$$

As mentioned in previous sections, the form of Eqn. 5 may differ depending on the implementing organization and the project under focus. We assume the above equation as a proof of concept of the generalizability of our algorithm. Needless to say, the above equation yields the maximum confidence level when all three weights, a_i , b_i and c_i are zero. However, this is often not a feasible condition and a practical value of the same often lies in between the bounds $a_i = (0, E_i)$, $b_i = (0, F_i)$ and $c_i = (0, M_i)$. The above equation maybe optimized using PSO for a particular set of values of a_i , b_i and c_i to obtain the idealized value for C. A snapshot of the simulation for $a_i = 2$, $b_i = 3$ and $c_i = 4$ (sample estimates for our simulation) is shown in Fig. 1 for iteration number 50, which proves the convergence of the PSO to an optimum value for the above equation. The two sub-figures show the F_i vs. E_i and the M_i vs. E_i plots as separate 2D convergences.

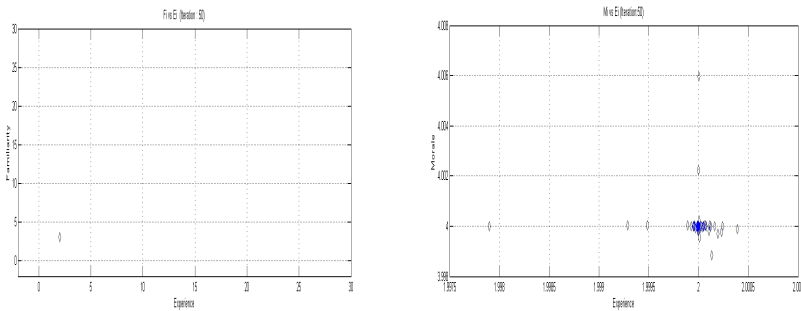


Fig. 1. Confidence attribute metrics optimized through PSO (plots show convergence after iteration number 50)

The efficiency of the PSO approach becomes apparent from the speed of convergence of Eqn. 5. The value of C converges rapidly to an optimum as shown in Fig. 2 to a relative value of 0 as per Eqn. 5. The speed and accuracy of the convergence is found to be better than that of the fuzzy approach [23].

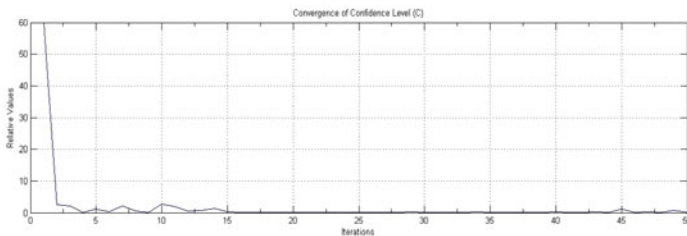


Fig. 2. Convergence of relative confidence levels

From Fig. 2, it is clear that the value of C converges by iteration 10 and remains relatively stable henceforth. This way of reliably estimating the value of C over a large number of iterations, is an important contribution of the PSO approach. We have

also evaluated this approach using multiple different values of a_i , b_i and c_i , as well as with higher order equations, and the performance remains within acceptable deviations of the above.

Once the value of C is decided from Eqn. 5, the updated KLOC value can be calculated using Step 5 of the algorithm. The SDE is calculated using COCOMO as usual using Eqn. 1. Now, an equation for P can be written as a combination of the TEDs in a similar fashion as Eqn. 5. Here too, we assume a condition where the various TEDs influencing our fictitious software development project are SQ , SC , and $WFDs$. While SQ and SC impact the STE positively in the sense that an increase in any of these metrics lead to a rise in STE, the $WFDs$ are inversely proportional to STE. Hence, the equation for P may be framed as a product of these drivers as follows:

$$P = (SQ - x_i) * (SC - y_i) * 1/(WFD - z_i) \tag{6}$$

Similar to Eqn. 5, here also, the x_i , y_i and z_i constants are appropriately selected based on the organizational and project nature, and the criticality that is attributed to each of the TEDs. Similar to the above case, a snapshot of the simulation for $x_i = 2$, $y_i = 3$ and $z_i = 4$ is shown in Fig. 3 for iteration number 50, which proves the convergence of PSO to an optimum value for Eqn. 6. The two sub-figures show SC vs. SQ and WFD vs. SQ as separate 2D convergences.

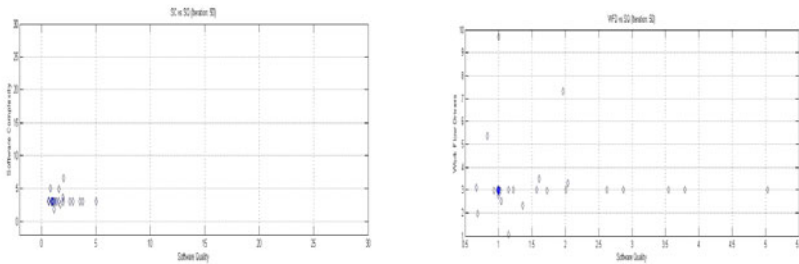


Fig. 3. TEDs optimized through PSO (plots show convergence after iteration number 50)

As with the equation for C , the P value can also be optimized efficiently using the PSO. The P value converges rapidly to an optimum as shown in Fig. 4 to a relative value of 0, as per Eqn. 6.

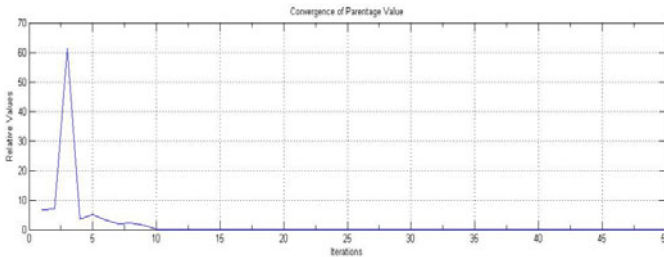


Fig. 4. Convergence of relative parentage values

Once the value of P in Eqn. 6 is optimized using PSO, the STE can easily be calculated as the product of the SDE and P (see Step 9 of the algorithm), which would give us a better STE approximation of the software development project under the given conditions of C and TEDs.

8 Conclusions and Future Scope

In this paper, we have presented and demonstrated an extension of the popular COCOMO approach for SDE estimation using PSO. The next step in this on-going research is to evaluate the performance of this methodology on live industry-grade software engineering projects and compare this approach to other soft computing alternatives. At present, it is our firm hypothesis that the proposed approach would be more effective than existing heuristic alternatives.

References

1. Nau, P., Randell, B.: Software engineering: Report of a Conference Sponsored by the NATO Science Committee. Scientific Affairs Division, NATO, Garmisch (1968)
2. Jalote, P.: An Intergraded Approach to Software Engineering, 2nd edn. Narosa Publishing House, India (2006)
3. Myers, G.J.: The Art of Software Testing, 1st edn. John Wiley and Sons, USA (1979)
4. Harrold, M.J.: Testing: A Roadmap, Proc. of 22nd International Conference on Software Engineering. In: Future of Software Engineering Track, Limerick, Ireland (2000)
5. Beizer, B.: Software Testing Techniques, 2nd edn. Van Nostrand Reinhold Company Limited, UK (1990)
6. Nageswaran, S.: Test Effort Estimation using Use Case Points, Quality Week, San Francisco, California, USA (2001)
7. Glass, R.L., Collard, R., Bertolino, A., Bach, J., Kaner, C.: Software Testing and Industry Needs. *IEEE Software* 23(4), 55–57 (2006)
8. Boehm, B.W.: Software Engineering Economics. Prentice-Hall, USA (1981)
9. Pressman, R.S.: Software Engineering: A Practitioner's Approach, 6th edn. McGraw-Hill, USA (2004)
10. Somerville, I.: Software Engineering, 7th edn. Pearson Education, India (2005)
11. Boehm, B.W., Abts, C., Brown, A.W., Chulani, S., Clark, B.K., Horowitz, E., Madachy, R., Reifer, D., Steece, B.: Software Cost Estimation with COCOMO II. Prentice-Hall, USA (2000)
12. Rubin, H.: Worldwide Benchmark Project Report. Rubin Systems Inc. (1995)
13. Hamer, P.G., Frewin, G.D.: M.H. Halstead's Software Science - A Critical Examination. In: Proc. of International Conference on Software Engineering, Tokyo, Japan, pp. 197–206 (1982)
14. Halstead, M.H.: Software Science-A progress report. In: Second Software Life Cycle Management Workshop, Atlanta, GA, USA (1978)
15. Kushwaha, D.S., Misra, A.K.: Software Test Effort Estimation. *ACM SIGSOFT Software Engineering Notes* 33(3) (2008)
16. Aranha, E., Borba, P.: An Estimation Model for Test Execution Effort. In: Proc. of 1st International Symposium on Empirical Software Engineering and Measurement, Madrid, Spain, pp. 107–116 (2007)

17. Jorgensen, M.: Realism in Assessment of Effort Estimation Uncertainty: It Matters How You Ask. *IEEE Transactions on Software Engineering* 30(4), 209–217 (2004)
18. Dawson, C.W.: An Artificial Neural Network Approach to Software Testing Effort Estimation. In: *Information and Communication Technologies*, vol. 20, Transaction of the Wessex Institute, UK (1998)
19. Srivastava, P.R., Saggur, S., Singh, A.P., Raghurama, G.: Optimization of Software Testing Effort using Fuzzy Logic. *International journal of Computer Sciences and Engineering Systems* 3(3), 179–184 (2009)
20. Srivastava, P.R.: Estimation of Software Testing Effort: An Intelligent Approach. In: *Accepted at Proc. of 20th International Symposium on Software Reliability Engineering (IEEE ISSRE 2009)*, Bangalore, India (2009)
21. Kennedy, J., Eberhart, R.C.: Particle Swarm Optimization. In: *Proceedings of IEEE International Conference on Neural Networks*, Piscataway, NJ, pp. 1942–1948 (1995)
22. Eberhart, R.C., Kennedy, J.: A New Optimizer using Particle Swarm Theory. In: *Proceedings of the Sixth International Symposium on Micro-machine and Human Science*, Nagoya, Japan, pp. 39–43 (1995)
23. Sheta, A., Rine, D., Ayesh, A.: Development of Software Effort and Schedule Estimation Models Using Soft Computing Techniques. In: *IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence)*, Hong Kong, pp. 1283–1289 (2008)

Prediction of E.coli Protein-Protein Interaction Sites Using Inter-Residue Distances and High-Quality-Index Features

Brijesh Kumar Sriwastava¹, Subhadip Basu^{2,*},
Ujjwal Maulik², and Dariusz Plewczynski³

¹ Department of Computer Science and Engineering,
Government College of Engineering and Leather Technology,
Kolkata-700098, India

² Department of Computer Science and Engineering,
Jadavpur University, Kolkata-700032, India

³ Interdisciplinary Centre for Mathematical and Computational Modeling,
University of Warsaw, 5a Street, 02-106 Warsaw, Poland
subhadip@cse.jdvu.ac.in

Abstract. We propose computational method for identification of protein-protein interaction sites using sequence and structure information. The method is trained on database of interacting proteins (DIP) for E.coli. Proteins that are known to interact are first collected from experimental results. All interacting partners are mapped onto corresponding three-dimensional structures. The training dataset for support vector machine algorithm is trained using both sequence composition and structural conformations of selected structures, if and only if both partners are composing the same complex. Our computational method is able to predict interactions for E.coli with 0.93 AUC, 0.89 sensitivity and 0.98 specificity.

1 Introduction

Protein–protein interactions (PPIs) are critical for understanding biological processes taking place in living cells. In order to understand the physico-chemical mode of binding typically the interaction site is analyzed. The identification of structural motifs observed in PPIs, responsible for molecular recognition process provides the important contribution for further automatic methods. The prediction of protein interactions networks (PINs), metabolic and signal transduction networks are of crucial importance for modern drug design [1]. The proteins used in these processes are extremely diverse, but their recognition sites have mostly same common properties. Actually, interaction sites have specific chemical and physical characteristics, all of which contribute to the molecular recognition process and these sites have been observed to be hydrophobic, planar, globular and protruding.

Currently developed experimental methods, such as yeast two-hybrid, or mass spectrometry applied to obtain protein-protein interactions (PPIs) have exposed the

* Corresponding author.

global view of the interaction network [2-6]. Moreover, in the context of drug development, the detection of functional modules is crucial for understanding cellular organization, and its intrinsic dynamics [7]. Typically above binary identification of interacting partners, and their sequence X-ray crystallographic or NMR experiments can enrich biological information providing the atomic-level structural details of protein-protein complexes. The binding conformation and interface physico-chemical characteristics are essential for understanding biological function of biomolecule, and design successful inhibitors that are used as drugs..

Accessible surface area measures are of great importance for detection of interacting partners [8]. The protein-protein interfaces are linked with the change in their solvent ASA, when going from monomeric to dimeric state. Typically, interface residue is identified, where ASA is decreased by 1Å. Permanent complexes have protein-protein interfaces, which are more closely packed but less planar, and with fewer inter sub-unit hydrogen bonds, when compared with the nonobligatory complexes [8].

Presently, a lot of effort is focused on detailed characterization of interface residues. In the current work, we use subset of physico-chemical features selected by consensus fuzzy clustering technique from a large set of 544 indices of AAindex1 database (<http://www.genome.jp/aaindex/>) [9]. The high quality indices HQI datasets are especially powerful for analyzing functional motifs in protein sequences by clustering, or machine learning techniques [10]. Deng et al. proposed the ensemble method, which combines bootstrap sampling technique, SVM-based fusion classifiers and weighted voting strategy to effectively utilize a wide variety of heterogeneous features [11].

Our paper solves the interaction problem; “given the unbound structures of two proteins predict their interface residues”. The information about their sequences, complexes with other interacting partners, and active sites description is used to statistically model the interfaces. More specifically, we are using sliding window of 21 amino acids to characterize the local sequence-structure interaction motifs,. We selected E.coli as the model organism for our study, where interaction partners are analyzed, and site prediction performance is evaluated by SVM classifier [12].

2 Materials and Methods

There are several sources of biological information on protein sequences, structures and their interactions are available online and we can divide these resources into two groups: sequence and structural. The first group of experimentally confirmed protein-protein pairings involves transient interactions, and the second focuses on complexes, i.e. stable interactions. In almost all of the databases, the developers use their own format for the data, making the integration across different datasets difficult. However theoretical analysis of interactions depends on heterogeneous sources of biological information, such as sequence and structural databases, the literature, and experimental data. The main databases containing experimental information about protein-protein interactions, which we have used, are: the Database of Interacting Proteins (DIP, <http://dip.doe-mbi.ucla.edu/dip/>)[13] and the Protein Data Bank (PDB) [14] database where one can find the three-dimensional structures of protein complexes.

Initially we started with 12606 number of protein-protein interactions of E.coli organism which are given in the file Ecoli20100614.txt of DIP database. Among these interactions, some entries did not have uniprotkb-id, matching PDB-id or even the primary sequences. After processing them they reduced to 2255 interactions. Again these interactions are verified for availability of the same PDB entry for both interacting protein, and then the size gets reduced to 312 entries. Each such entry now comprises of a valid PDB-id (for the protein-protein complex), with multiple uniprotkb-ids. Now the entries for homo protein interactions are also removed and we get only 40 valid hetro interactions, for our E.coli database. We got the amino acid sequences from file dip20091230.seq (available at the DIP web server) using the corresponding uniprotkb-ids. The structured database format, used for our work is shown below:

```
{id Protein A},{id Protein B},{Organism Name},{Interaction Type},{PubMed id},{Database Name},{PDB id A},{PDB id B},{lenght of amino acid A},{amino acid sequence of Protein A},{lenght of amino acid B},{amino acid sequence of Protein B}
```

2.1 Design of Feature Set

In conjunction with machine and statistical learning approaches, we performed an extensive search to derive, optimize, and evaluate features that can best discriminate between interacting and non-interacting sites. These features can be roughly divided into eight groups which are Electric properties, Hydrophobicity, Alpha and turn propensities, Physicochemical properties, Residue propensity, Composition, Beta propensity and Intrinsic propensities. Currently, 544 amino acid indices are released in AAindex1 database. These features were clustered into different High-Quality-Indices (HQI) by Saha et al. [9]. In the current work we have used 8 HQIs (HQI8) described as, BLAM930101, BIOV880101, MAXF760101, TSAJ990101, NAKH920108, CEDJ970104, LIFS790101 and MIYS990104. Detail description of the clustering method, software and supplementary material are given at <http://sysbio.icm.edu.pl/aaindex/AAindex/>.

2.2 Feature Representation

In this work we are working with interacting protein pairs (say, P_A and P_B), as described in our aforementioned database. Now let P_A and P_B has their own amino acid sequences as a_1, a_2, \dots, a_M and b_1, b_2, \dots, b_N respectively, where $a_i, b_j \in \{A, R, N, D, L, K, M, F, C, Q, E, G, H, I, P, S, T, W, Y, V\}, \forall i = 1 \text{ to } M \text{ and } \forall j = 1 \text{ to } N$.

Now we compute inter-atom distances between P_A and P_B . Please note that we consider only the heavy atoms (as given in respective PDB entry) from each amino acid for this purpose. We define the distance measures as follows:

$$D_P(a_i, b_j) = \min(d_r(a_{ik}, b_{jl})), \forall k = 1 \text{ to } P \text{ and } \forall l = 1 \text{ to } Q,$$

where, P and Q are number of heavy atoms in the residues a_i and b_j respectively and $d_r(a_{ik}, b_{jl})$ = inter-atom Euclidean distances between the k^{th} heavy atom of a_i and l^{th} heavy atom of b_j .

Now when $D_p(a_i, b_j)$ is lower than 3.5 \AA , then corresponding residue pair (a_i, b_j) corresponding to the protein pair (P_A, P_B) is said to be interacting, otherwise they are said to be non-interacting.

The protein sequences are now virtually fragmented into multiple overlapping sub-sequences each consisting of 21 amino acids. Now for the proteins P_A and P_B we consider the residues from a_1, a_2, \dots, a_{21} and b_1, b_2, \dots, b_{21} respectively, and check whether any of the residue pairs have $D_p(a_i, b_j) < 3.5 \text{ \AA}$. If found, we annotate the pair of sub-sequences (obtained from P_A and P_B respectively) as one positive interaction and extract HQI8 features for the 42 residues, resulting in a $42 \times 8 = 336$ dimension positive feature vector. The overlapping subsequences are then shifted, like a sliding window, to check for further positive interactions. In all cases, where two sub-sequences have no interacting residue pair, then such sub-sequence pair is said to be non-interacting and we extract negative 336 features values for it using HQI8 features. These feature values are then used by the machine learning procedure to train/test the support vector machines, designed separately to produce optimal recall, precision and AUC (Area Under ROC curve) scores.

3 Experimental Result

As discussed before, we have prepared interacting and non-interacting residue fragments and extracted HQI8 features for both positive and negative data samples. In all we found 1701 positive interactions, and from all negative interactions randomly chose 3213 data samples. This dataset is then partitioned into mutually exclusive training and test sets in the ratio 8:1. These two data clusters finally reduce to be a problem of binary classification, which are handled by a nonlinear support vector machine with polynomial kernel function of degree 5. Training is performed on three different optimizing criterion, viz., recall, precision and AUC score. During training we performed 3-fold cross validation and using all the three training networks, we evaluate the test dataset. These three experiments are marked as run#1, run#2 and run#3 respectively. Figure 1 shows results for 3-fold cross-validation experiment on the train and the test data (average and maximum performances are also shown), for the AUC optimized network. In the current work we present results over the test data (for all the three runs) using Recall, Precision and AUC optimized networks (see Tables 1-3). We also design a quality consensus scheme that predicts an interaction to be positive when one, two or all three aforementioned optimized networks decide an interaction to be positive (see result in Table 4).

Now we compare the current findings with similar works reported in the literature. In the work of Wang et al [15] position specific scoring matrices (PSSMs) were used along with evolutionary conservation score for 11 neighbor residues. They obtained 71.9% ASC, 68.6 % Sensitivity and 65.4% Specificity over their PPI dataset. Nguyen et al. [16] used PSSMs and accessible surface areas (ASA) with 15 neighbor residue to get 74.9% AUC, 35.9% Sensitivity and 92.9% Specificity scores. Both of them used SVM to construct classifier. Deng et al [11] uses an ensemble method with weighted voting strategy along with SVM approach and achieved 79.7% AUC, 76.7% Sensitivity and 63.1% Specificity. Borderner et al. [17] achieved 76% Accuracy, 57% Recall and 26% Precision. Rohit Singh et al. [18] obtained 60% Sensitivity and 75%

Specificity. In comparison, we obtained 93.26% AUC (obtained in run#1 of consensus prediction, see Table 4), 88.7574% Sensitivity (or Recall, see Table 4) and 97.76% Precision (see Table 4) over our E.coli test dataset. We have also added a comparison table (see Table 5) and an illustration (see Figure 2) for easy reference.

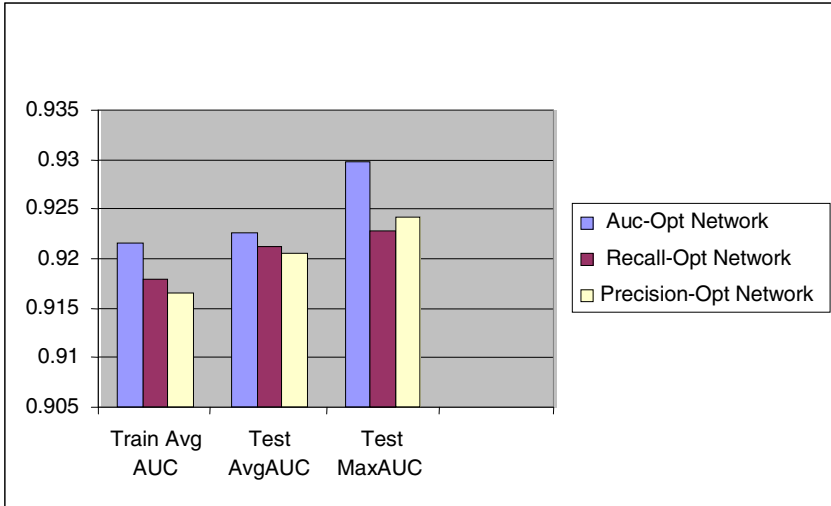


Fig. 1. Results for 3-fold cross-validation experiment for the AUC optimized network

Table 1. Result on AUC optimized network over E.coli test data

Run	Accuracy	Recall	Precision	Specificity	AUC
run#1	94.48669	0.887574	0.9375	0.971989	0.929781
run#2	93.34601	0.881657	0.908537	0.957983	0.91982
run#3	93.15589	0.881657	0.90303	0.955182	0.918419

Table 2. Result on Recall optimized network over E.coli test data

Run	Accuracy	Recall	Precision	Specificity	AUC
run#1	84.79088	0.887574	0.7109	0.829132	0.858353
run#2	93.53612	0.887574	0.909091	0.957983	0.922779
run#3	93.72624	0.881657	0.919753	0.963585	0.922621

Table 3. Result on Precision optimized network over E.coli test data

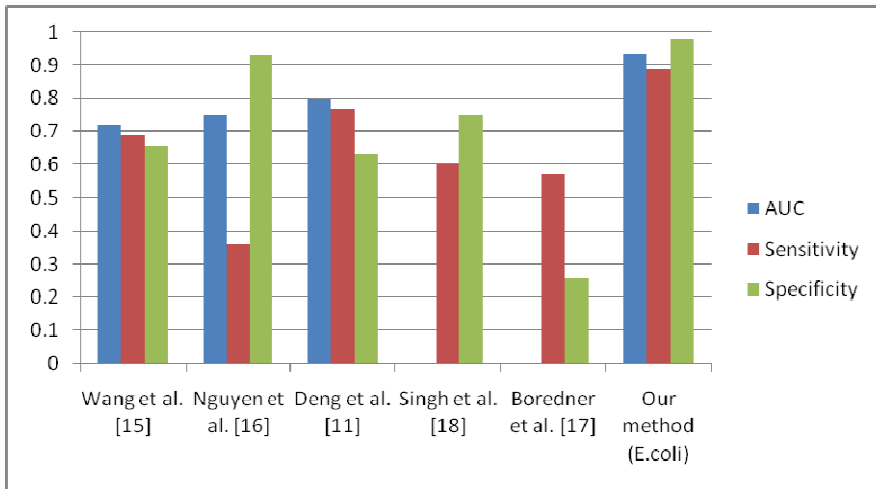
Run	Accuracy	Recall	Precision	Specificity	AUC
run#1	92.96578	0.887574	0.892857	0.94958	0.918577
run#2	92.96578	0.887574	0.892857	0.94958	0.918577
run#3	93.72624	0.887574	0.914634	0.960784	0.924179

Table 4. Result of consensus prediction over E.coli data

Run	Accuracy	Recall	Precision	Specificity	AUC
run#1	94.8669	0.887574	0.949367	0.977591	0.932583
run#2	93.7262	0.887574	0.914634	0.960784	0.924179
run#3	93.1559	0.887574	0.898204	0.952381	0.919977

Table 5. A comparative analysis with similar works available in the literature

Methods	AUC	Sensitivity	Specificity
Wang et al. [15]	0.71933	0.68640	0.65417
Nguyen et al. [16]	0.74943	0.3598	0.92949
Deng et al. [11]	0.79761	0.76765	0.63158
Singh et al. [18]	-	0.6	0.75
Boredner et al. [17]	-	0.57	0.26
Our method (E.coli)	0.932583	0.887574	0.977591

**Fig. 2.** An illustrative comparison of our work with similar works available in literature

4 Conclusion

Protein-protein interactions are important for enriching significant biological knowledge. They can be used as the starting point for understanding, how the cell works internally, as a collection of biomolecules. In this study we propose the method for PPIs prediction using both protein sequence and its three-dimensional structure. The distance between all atom pairs from interacting proteins is calculated, if it is equal or less than 3.5\AA this pair is considered as interacting ones. The local sequence neighborhoods are collected and HQI features vectors are used to represent the

contiguous, overlapping sliding window of length 21 residues. Finally, support vector machine algorithm, with polynomial function of the degree 5 is used to build statistical learning model. This model can be further tested as predictor, which allow to annotate unknown interactions, enriching the biological knowledge about proteins partners. Our classification results are better than Rohit Singh et al. [18] obtained. We conclude that our machine learning method combined with feature selection algorithm that utilize HQI8 indices [9] is able successfully predict the interaction residues with small error rate (below 5%).

Acknowledgements. This work was supported by the Polish Ministry of Education and Science (grant: N301 159735) and Department of Science and Technology, India (Grant No. DST/INT/MEX/RPO-04/2008(ii)). Contribution of second author is also supported by the PURSE project of Computer Science and Engineering Department of Jadavpur University, India.

References

1. Chelliah, V., Chen, L., Blundell, T.L., Lovell, S.C.: Distinguishing structural and functional restraints in evolution in order to identify interaction sites. *J. Mol. Biol.* 342, 1487–1504 (2004), doi:10.1016/j.jmb.2004.08.022
2. Uetz, P., Giot, L., Cagney, G., et al.: A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* 403, 623–627 (2000)
3. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., Sakaki, Y.: A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* 98, 4569 (2001)
4. Gavin, A.C., Bösch, M., et al.: Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141–147 (2002)
5. Ho, Y., Gruhler, A., Heilbut, A., et al.: Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415, 180–183 (2002)
6. Gavin, A.C., Aloy, P., Grandi, P., et al.: Proteome survey reveals modularity of the yeast cell machinery. *Nature* 440, 631–636 (2006)
7. Krogan, N.J., Cagney, G., Yu, H., Zhong, G., et al.: Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* 440, 637–643 (2006)
8. Jones, S., Thornton, J.: Principles of protein–protein interactions. *Proc. Natl. Acad. Sci. USA* 93, 13–20 (1996)
9. Saha, I., Maulik, U., Bandyopadhyay, S., Plewczynski, D.: Fuzzy Clustering of Physicochemical and Biochemical Properties of Amino Acids. *Amino Acids* (accepted)
10. Deng, L., Guan, J., Dong, Q., Zhou, S.: Prediction of protein–protein interaction sites using an ensemble method. *BMC Bioinformatics* 10, 426 (2009), doi:10.1186/1471-2105-10-426
11. Deng, L., Guan, J., Dong, Q., Zhou, S.: Prediction of protein–protein interaction sites using an ensemble method. *BMC Bioinformatics* 10, 426 (2009), doi:10.1186/1471-2105-10-426
12. Vapnik, V.: *The Nature of Statistical Learning Theory*. Springer, New York (1995)
13. Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., Eisenberg, D.: The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* 32, D449–D451 (2004)
14. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The Protein Data Bank. *Nucleic Acids Res.* 28, 235–242 (2000)

15. Wang, B., Chen, P., Huang, D.S., Li, J.J., Lok, T.M., Lyu, M.R.: Predicting protein interaction sites from residue spatial sequence profile and evolution rate. *FEBS Letters* 580, 380–384 (2006)
16. Nguyen, M.N., Rajapakse, J.C.: Protein-Protein Interface Residue Prediction with SVM Using Evolutionary Profiles and Accessible Surface Areas. In: *CIBCB*, pp. 1–5 (2006)
17. Bordner, A.J., Abagyan, R.: Statistical Analysis and Prediction of Protein–Protein Interfaces. *PROTEINS: Structure, Function, and Bioinformatics* 60, 353–366 (2005)
18. Singh, R., Park, D., Xu, J., Hosur, R., Berger, B.: Struct2Net: a web service to predict protein–protein interactions using a structure-based approach. *Nucleic Acids Research* 38, W508–W515 (2010) (web server issue)

More Secured Text Transmission with Dual Phase Message Morphing Algorithm

M. James Stephen¹, P.V.G.D. Prasad Reddy², Ch. Demudu Naidu³,
Sampangi Sonali⁴, and Ch. Heymaraju⁵

¹ Department of I.T, ANITS, Visakhapatnam, India
jamesstephenm@yahoo.com

² Dept. of CS & SE, Andhra University, Visakhapatnam, India
prasadreddy.vizag@gmail.com

³ Department of I.T, ANITS, Visakhapatnam, India
naidu.061@gmail.com

⁴ Dept. of I.T, ANITS

⁵ Dept. of I.T, GIT, G.U

Abstract. In this global village, where with the advancements in the field of communications, ensuring security to the data being transmitted has become very vital. These safety and security issues contributed to the outgrowth of secret communication. While encryption simply encodes the data making it difficult for the layman to understand, Steganography deals with hiding data within another data making them unaware of it. Both concepts ensure data security but in different forms. These two approaches when unified provide much better security to the data than that provided by either of encryption or Steganography.

This paper proposes such a new unified approach for Secured Text Transmission using Dual Phase Message Morphing (DPMMA) algorithm, which encrypts and conceals data in two consecutive stages to provide better security to the data. This algorithm is a simple unified approach of encryption and Steganography which employs two newly proposed techniques for encryption and Steganography to provide better security for the data.

As the name suggests it works in two consecutive phases. In the first phase encryption is performed and in the second phase the encrypted message is concealed within another text. The result of these two phases produces a morphed text which does not resemble the original message.

Keywords: E-Message, Message, Cover Message, Sequence, Label, Encryption, Text Steganography.

1 Introduction

The advent of internet has benefited the human life up to a greatest extent by connecting the entire globe on a single medium. E-mails, chat rooms, social networking sites, data sharing etc all these features catalyzed the widespread of internet. Many attackers try to intrude into the network and capture the data being transmitted. Thus providing security to data being transmitted has turned out to be a

crucial task during communications. The best opted methods for secret communications are Data Encryption and Steganography. Steganography is the art and science of hidden writing. While an encryption program protects your message from being read by those not in possessions of the key, sometimes you wish to obscure the very fact you're sending an encrypted message at all.

The comparison of different techniques for communicating in secret can be found in [1]. The unified approach of the above two techniques will provide a better security. Several existing Steganography methods can encrypt data before hiding it in the chosen medium [2]. But all these methods use image as their cover data, where as the present method employees text as the cover data.

2 Why Text Steganography?

Text Steganography deals with hiding information in simple text data but not digital data. It is relatively easy method to hide data and makes the process of breaking the hidden message difficult unless technique used is known. The comparison of various forms of Steganography is as shown in the Table 2 [3].

Table 1. Comparision of Various Forms of Steganography

Steganography Techniques	Medium	Embedding Technique
Steganography Using Text	Text files	To embed information we need to simply alter the text to a suitable form.
Steganography Using Audio	MP3 files	Encode data as a binary sequence which sounds like noise
Steganography Using Image	Image files	It works by altering the bit configurations or wavelets.
Steganography Using Video	Video files	A combination of sound and image techniques can be used.

From the above table we can see that text Steganography is the simplest form of Steganography. The best thing about this text Steganography is that the hidden information cannot be extracted unless the mode of embedding is known.

There are number of techniques existing techniques for text Steganography [4] like ‘Steganography of Information in Random Character and Word Sequences’ , ‘Steganography of Information in Specific Characters in Words’ ,’Creating Spam Texts’, ‘Line Shifting’, ‘Word Shifting’ , ‘Syntactic Methods’ , ‘Semantic Methods’, ‘Feature Coding’, ‘Abbreviation’, ‘Open Spaces’, ‘Persian/Arabic Text Steganography’

Rather than placing the alphabet in particular positions, it would be a better idea to place the characters at desired random positions making it difficult for the attacker to break the hidden message. Hence the message is secured. What if the message is

encrypted before placing it in random positions? Then the message becomes doubly secured. This particular point works as the basic principle for our suggested DPMMA algorithm.

3 Proposed Work

Here we present a new method for text Steganography using DPMMA (Dual Phase Message Morphing Algorithm). This algorithm is a simple unified approach which employs two newly proposed techniques for encryption and Steganography to provide better security for the data.

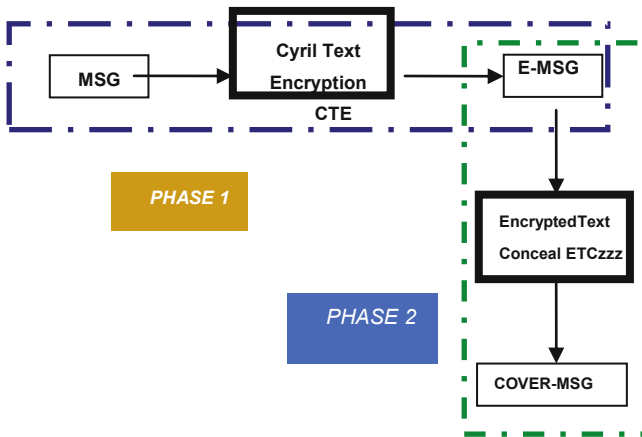


Fig. 1. Structure of DPMMA

As the name suggests it works in two consecutive phases. In the first phase encryption is performed and in the second phase the encrypted message is concealed within another text. The result of these two phases produces a morphed text which doesn't resemble the original message.

3.1 First Phase of DPMMA

The first stage performs encryption. The original text (Message) is encrypted to produce the cipher text (E -Message). Encryption is the method of changing the text from its original form to another form that cannot be easily understood. Substitution techniques are the simplest encryption techniques. In this algorithm a new substitution technique named CTE (Cyril Text Encryption) is introduced.

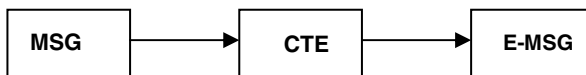


Fig. 2. First phase of DPMMA

CTE is a substitution technique based on a table called Cyril Tab. This table is generated based on a Russian font called Cyrillic font. The character in the original text (Message) get substituted by another character depending on the table thus producing the encrypted text (E-Message).

Table 2. CYRIL TAB

a	b	c	d	E	f	g	h	i	J	k	l	m
f	u	c	t	Y	a	n	p	i	O	z	d	b
n	o	p	q	R	s	t	u	v	W	x	y	z
m	w	x	q	K	g	e	s	v	J	l	h	r

Special characters and numeric data if used in message are not encrypted but remain the same in e-message also. Based on the above table the substitution is performed and the E-Message is now passed to next phase to create Cover message.

Algorithm for Encryption Using Cyril Tab.

```

a : message
e : e-message
start cte ( a )
for each character a[i] do
encrypt a[i] using cyril tab
end for
return e
end cte

```

Fig. 3. Pseudo code of CTE technique

3.2 Second Phase

Johnson and Katzenbeisser grouped steganographic techniques into six categories depending on how the algorithm encodes information in the cover object. They are: substitution systems, transform domain techniques, spread spectrum techniques, statistical methods, distortion techniques, and cover generation methods[5]. Cover generation techniques are most unique of these six types. A cover generation method actually creates a cover for the sole purpose of hiding information.

This phase deals with concealing the E-Message generated from the first phase. To conceal the E-Message we generate another text called the Text Passage using ETC (Encrypted Text Conceal) technique which is a new cover generation steganography technique. This technique involves two important features, one is Text Passage and the other is Label which will be seen later.

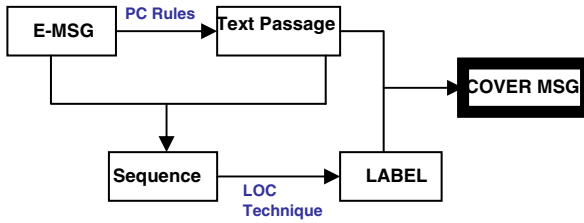


Fig. 4. Second phase of DPMMA

To generate the Text Passage from the E-Message certain rules are defined called as PC (Passage Creation) Rules.

- Rule 1:** Each alphabet in the E-Message word should correspond to each word in the Text Passage.
- Rule 2:** Each word in the E-Message should correspond to each sentence in the Text Passage.
- Rule 3:** Every alphabet in the E- Message can be placed in any desired position (between 1 to 9) in the words in the Text Passage.
- Rule 4:** Every sentence in the text passage should end with Full stop and unwanted spaces should be avoided.

Algorithm for labeling the cover message

```

e : e-message
p : text passage
s: sequence
start etc( e , p)
w=split p to array of strings;
for each e[i] do
    s[i]=index of e[i] in s[i]
end for
label=loctechnique( s[i] )
add label to p
end etc
  
```

Fig. 5. ETC technique

Based on the above rules the text passage is generated. It is made clear that every word in the passage contributes to the hidden message. At this point it is understood that hidden message is embedded within the passage at random positions (between 1 to 9) as desired by the user. The positions at which the data in the E-Message is placed, when combined together produce a Sequence. This sequence is actually generated from the E-Message and the Text Passage. Sequence plays the key role in extracting the hidden data from the passage. Hence this sequence also needs to be secured.

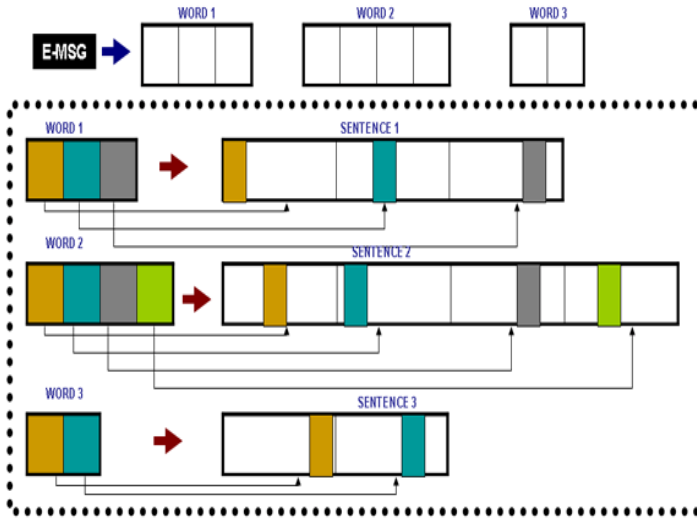


Fig. 6. PC Rules

(Long Octal conversions) technique. This technique is an encryption technique used for the encryption of numeric data. In this technique the sequence is initially converted to a set of long numbers which are then converted to their corresponding octal strings and finally then concatenated to produce a string called Label.

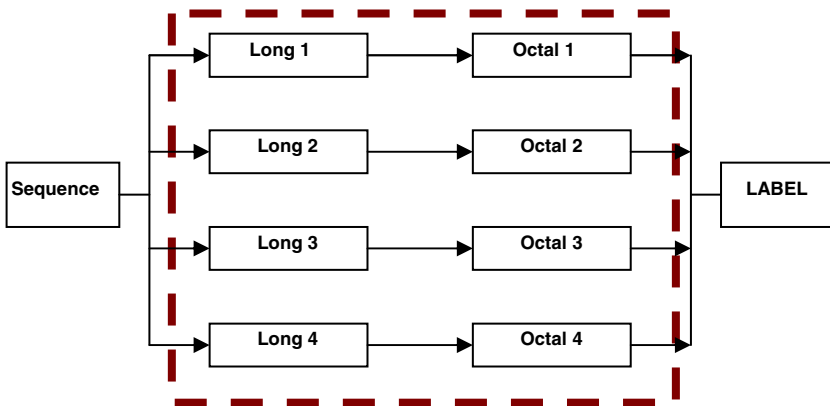


Fig. 7. LOC Technique

This Label is then appended at the end of the text passage to create the cover data (Cover Message) which can then be sent to the required person. This Label concept is analogous to digital signature but method of computation is totally different. This completes the generation of cover message using ETC technique. Now this cover message which is the required cover data can be used for secret communications.

Algorithm for LOC technique

```

start loctechnique( s )
initialize k=i+19 and label,octal as null;
for each i less than equal to s.length do
l[j]=convert s[i] to s[k] as long and adjust I,j,k
end for
for each l[i] do
    o[i]=octal string of l[i]
    octal=add octal and o[i]
end for
return octal
end loctechnique

```

Fig. 8. LOC technique

Extracting the data from the cover message can be performed in the reverse process of the employed techniques. The label is first extracted from the cover message and then it is decrypted to produce the sequence. The sequence is then used to extract characters from the passage. The extracted data when decrypted produce the original hidden message.

Algorithm for Extraction of Data from cover message

```

start extract ( p, s )
l=label,o= octal strings
for each i less than equal to o.length do
l[i]=convert o[i] to its equivalent long
end for
for each j less than l.length
s[k]=convert l[i] to set of integers
end for
w=split p to array of strings;
for each s[i] do
    a[i]=character at s[i] in w[i]
end for
decrypt a[i] using cyril tab
end extract

```

Fig. 9. Pseudo code of extraction**4 Results**

As discussed the algorithm works in two consecutive stages. The first stage performs encryption to produce encrypted message using CTE technique. The number of characters in the message and e-message remain same. So no data loss can occur.

The second stage uses cover generation Steganography technique called ETC to generate the cover message along with label. Each character in e-message should correspond to each word in passage, and label is generated only when the number of characters and words match, hence data cannot be lost.

Our experiments proved that this proposed system is 99 % successful in providing better .99.5 % attacks to break the message completely resulted in failure. Since the system was developed entirely using Java, it works on any platform that supports java.

5 Conclusion

Security and privacy issues enhanced the growth of secret communications. Many Steganography techniques evolved to transmit hidden messages and at the same time lot of analysis works were carried out to detect the presence of hidden messages which raised the need for more robust Steganography approaches.

This paper proposes a robust and secure method of Steganography using DPMMA which encrypts and then hides the data. The cover data can be transmitted over mails, chats, text files , contents of web pages etc, where mostly text is used.

6 Future Implementation

This is a flexible user-friendly application developed to provide robust and secure transmissions of hidden data. We look forward to add a word suggestion algorithm to make it much more user friendly. Because of it flexibility and efficiency it can be added as a feature in mail systems or as an add-on on the web browsers.

References

- [1] Cummins, J., Diskin, P., Lau, S., Parlett, R.: Steganography and digital watermarking. School of Computer Science, The University of Birmingham (2003)
- [2] Artz, D.: Digital Steganography: hiding data within data. IEEE Internet Computing (May-June, 2001)
- [3] Channalli, S., et al.: International Journal on Computer Science and Engineering 1(3) (2009)
- [4] Shirali Shahreza, M.: Text steganography in sms. In: International Conference on Convergence Information Technology (July 2007)
- [5] Bennett, K.: Linguistic Steganography: Survey, Analysis, and Robustness Concerns for Hiding Information in Text, Purdue University, CERIAS Tech. Report 2004-13 (2004)

Removal of False Minutiae with Fuzzy Rules from the Extracted Minutiae of Fingerprint Image

M. James Stephen¹, P.V.G.D. Prasad Reddy², Vadlamani Kartheek³,
Ch. Suresh⁴, and Suresh Chandra Satapathy⁵

¹ Department of I.T, ANITS, Visakhapatnam, India
jamesstephenm@yahoo.com

² Dept. of CS & SE, Andhra University, Visakhapatnam, India
prasadreddy.vizag@gmail.com

³ Department of I.T, ANITS, Visakhapatnam, India

⁴ Dept. of I.T, ANITS

⁵ Dept. of CSE, ANITS, India

Abstract. Human fingerprints are rich in details called minutiae, which can be used as identification marks for fingerprint verification. Minutiae are the two most prominent and well-accepted classes of fingerprint features arising from local ridge discontinuities: ridge endings and ridge bifurcations. In today's world minutia matching is most popular and modern technology for fingerprint matching. If there is enough minutia point in one fingerprint image that are corresponding to other fingerprint image, then it is most likely that both images are from the same finger print. In this paper, we proposed a complete system for minutiae extraction and removing the false minutiae from the extracted ones.

The main objective of this paper is developing a new idea for extracting minutiae points and removing the false minutiae by implementing some fuzzy rules. It comprises of various steps. It begins with the acquisition of the fingerprint image. This is followed by binarization ie, converting the gray image to binary image and then thinning ie, making the ridges just one pixel wide. Finally the minutiae points are extracted based on Tico and Kuosmanen[1] and the Crossing Number(CN) method. Then, among the extracted minutiae, false minutiae are removed with fuzzy rules. Thus our system could be a better pre-processing technique for authentication.

Keywords: Fingerprint, Minutiae, Ridge, Bifurcation, Binarization, Thinning, False minutiae.

1 Introduction

The recent advances of information technologies and the increasing requirements for security have led to a rapid development of automatic personal identification systems based on biometrics. Biometrics [4, 5] refers to accurately identifying an individual based on his or her distinctive physiological (e.g., fingerprints, face, retina, iris) or behavioral (e.g., gait, signature) characteristics.

A fingerprint is the pattern of ridges and valleys on the surface of a fingertip. A total of eight different types of local ridge/valley descriptions have been identified [2].

Instead, in accordance with the representation of fingerprints in the U.S. Federal Bureau of Investigation (FBI) [3], ridge endings and bifurcations, called minutiae, are taken as the distinctive features of the fingerprints.

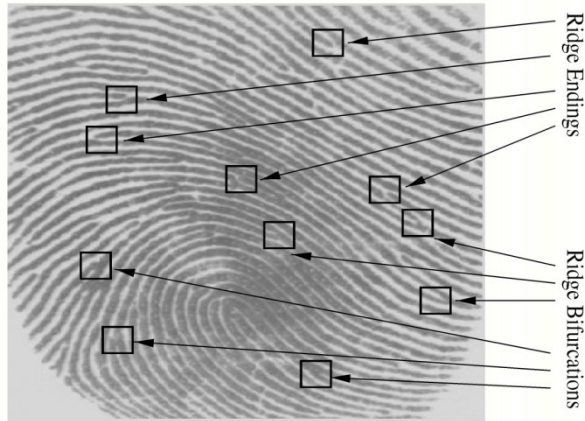


Fig. 1. Ridge endings and bifurcations

The method that is selected for fingerprint matching was first discovered by Sir Francis Galton. In 1888 he observed that fingerprints are rich in details also called minutiae in form of discontinuities in ridges. He also noticed that position of those minutiae doesn't change over the time. Therefore minutiae matching are a good way to establish if two fingerprints are from the same person or not.

Quality of fingerprints can significantly vary, mainly due to skin condition and pressure made by contact of fingertip on sensing device. This problem can be handled by applying an enhancing algorithm that is able to separate and highlight the ridges from background; this type of enhancing is also called binarization.

A more effective and faster minutiae extraction realization can be achieved by minimizing data that represents minutiae without corrupting it. Since minutiae are determined only by discontinuities in ridges, they are totally independent of ridges thickness. Thinning of the ridges to only 1-pixel wide lines also called skeletons, not only preserves minutiae but it does it with minimum possible data usage. The thinning method is often called skeletonization.

Most of the finger-scan technologies are based on Minutiae. Minutia based techniques represent the fingerprint by its local features, like terminations and bifurcations. This approach has been intensively studied, also is the backbone of the current available fingerprint recognition products [6].

Once the minutiae points are extracted from the thinned image, then the system proceeds to further task of removing the false minutiae. In the process of thinning, some points appear to be minutiae which actually are not. These false minutiae have to be removed. This is done by implementing some fuzzy rules and the actual minutiae points of the image could be stored.

2 Existing System

In the existing system of extracting the minutiae, the following are the various steps involved:

Step 1: Image Acquisition

The first step is acquiring the fingerprint image. This could be done in two ways online and offline. The former one is referred as “live-scan” image where as the latter is referred as “inked” fingerprint.

In the online process of acquiring the image, the fingerprint image is obtained instantly through a scanner and the operations are performed on that image.



Fig. 2. Fingerprint

In the other case, the stored images are acquired through various sources and the operations are performed on them.

Step 2: Image Binarization

Binarization is the process of converting gray image to binary image. The process of extraction of the minutiae is done on the binary image only. In this, there are only two levels of interest 0 for black level and 1 for white level. Once the operation is performed, ridges are highlighted with black color and valleys with white color.



Fig. 3. Binarised image

This process involves examining the grey level value of each pixel in the image and if the value is greater than the threshold, then the pixel value is set to a binary value of 1 or else 0. Finally, the outcome is a binary image containing two levels of information, the foreground ridges and background valleys.

A locally adaptive Binarization method is performed to binarize the fingerprint image. Such a named method comes from the mechanism of transforming a pixel value to 1 if the value is larger than the mean intensity value of the current block (16x16) to which the pixel belongs.

Step 3: Image Thinning

Skeletonization is a process mostly used on binarized images by thinning a certain pattern shape until it is represented by 1-pixel wide lines, the so called skeleton of that pattern.



Fig. 4. Thinned image

Since minutiae are determined only by discontinuities in ridges, they are totally independent of ridges thickness. By finding the skeleton of the binarized fingerprint image through thinning of the ridges to only one pixel wide lines, the minutiae are preserved with minimum possible data. This decimation of data offers more effective minutia extraction realization.

Thinning is the morphological process to remove the foreground pixel until they are one pixel wide. So in the first step morphological process apply to reduce the width of the ridge. There are two main morphological processes which are Erosion and Dilation. The former refers to thinning an object and the latter refers to the vice-versa. A thinning algorithm proposed by Guo and Hall is implemented[7].

Step 4: Minutiae extraction

Extraction minutiae point and their location is an important step of the process. There are many algorithms have been developed for minutiae point extraction. An algorithm based on Tico and Kuosmanen method [1] and Crossing Number (CN) methods is used for extract minutia points. This method extracts the ridge endings and bifurcations from the skeleton image by examining the local neighborhood of each ridge pixel using a 3x3 window. The CN for a ridge pixel P is given by [8]

$$CN=0.5\sum_{i=1}^8 |P_i - P_{i+1}| , P_9 = P_1 \tag{1}$$

Where P_i is the pixel value in the neighborhood of P. For a pixel P, its eight neighboring pixels are scanned in an anti-clockwise direction

After the CN for a ridge pixel has been computed, the pixel can then be classified according to the property of its CN value.

- If value of CN is one, then the central pixel is termination.
- If value of CN is two then the central pixel is usual pixel.
- If value of CN is three then the central pixel is bifurcation.

To find termination, first divided image into a image M of size WXW , then label 1 to all pixels in image M (3X3 image) which are eight connected with the termination point. After this step we count in clockwise direction, the number of 0 to 1 transition along to border. If the value of transition is equal to 1 then this minutia point will be consider as termination.

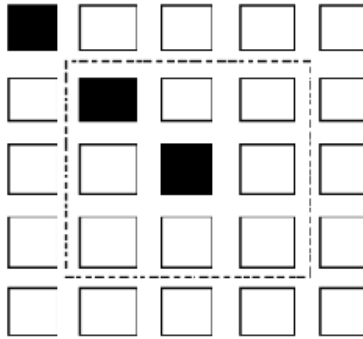


Fig. 5. CN=1 (Termination)

To find bifurcation first we examine the eight neighborhood pixels surrounding the bifurcation point in clockwise direction. Now label 1, 2 and 3 to the pixels that are connected to bifurcation point. After that, label to rest of ridge pixel that is connected to these three pixels, this labeling is similar to termination labeling.

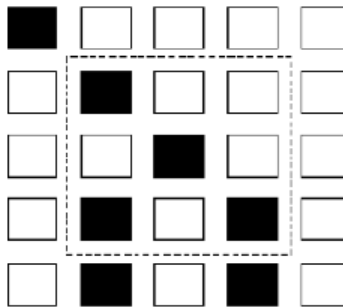


Fig. 6. CN=2 (Bifurcation)

After labeling we count in clockwise the no of 0 and 1, 0 to 2 and 0 to 3 transition along to the border of image. If the value of transition is equal to 1 then minutia point will be considered as bifurcation point.

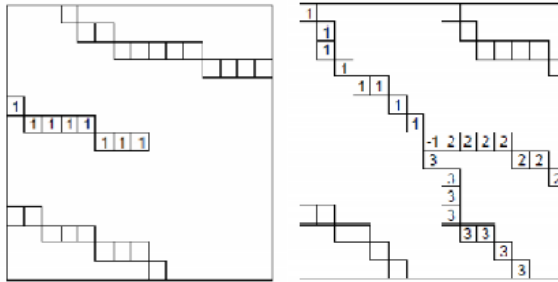


Fig. 7. Termination and Bifurcation

3 Proposed System

The proposed system is a step ahead of the existing system. Along with the various steps involved in the current system of extracting minutiae, another step of removing the false minutiae is added to the system for increasing the accuracy. Hence, the following are the various steps involved in our proposed system:

- Step 1:** Image Acquisition
- Step 2:** Image Binarization
- Step 3:** Image Thinning
- Step 4:** Minutiae extraction
- Step 5:** Removal of the false minutiae
- Step 6:** Exporting the minutiae to a text file

In order to find spurious point a new algorithm is proposed that is based on some fuzzy rules. This algorithm tests the validity of each minutiae point in thinned image and examines the local neighborhood around the point. The first step in this algorithm is to find the distance between termination and bifurcation. We have used Euclidian method to find distance [9]. After finding distance, we will use some rules to remove these false minutia points.

The proposed fuzzy rules are as follows:

- Rule1:** If the distance between termination and bifurcation is less than D , then remove this minutia.
- Rule 2:** If the distance between two bifurcations is less than D , then remove this minutia.
- Rule 3:** If the distance between two terminations is less than D , then remove this minutia.

After extraction of minutia, find location of these minutiae points. For finding minutiae points' location, the following three points have been obtained

- X and Y coordinates
- Orientation angle between these coordinates
- Type of minutiae (ridge ending or bifurcation)

Step 6: Exporting the minutiae to a text file

Once the false minutiae are removed, the minutiae points are exported to a text file in the workspace. The number of terminations and bifurcations could be compared in the cases of removing the false minutiae with the prior ones of non-removed ones so as to check the accuracy of the system in removing the false minutiae.

4 Experimental Results

Our system implemented the above mentioned steps in MATLAB 7.0 and obtained better results than the existing systems of extracting the minutiae. An offline image (FP img.1) is acquired from the workspace. Then the binarized and thinned images are obtained for this image.



Fig. 8. Binarized image



Fig. 9. Thinned image

From the thinned image, minutiae points are extracted with the Crossing Number approach.

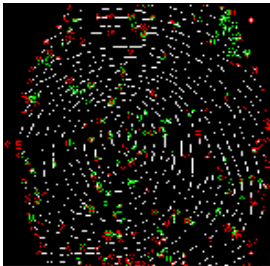


Fig. 10. Minutiae extraction

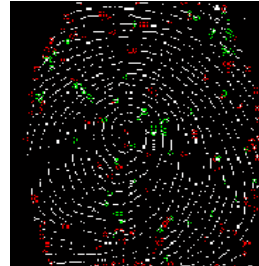


Fig. 11. Removal of false minutiae

The result is:

- No. Of Terminations: 234
- No. Of Bifurcations: 150

Now, the false minutiae are removed using the fuzzy rules.

The result after applying the Fuzzy rules is:

- No. Of Terminations: 124
- No. Of Bifurcations: 66

Thus, it could be observed that, there would be some false minutiae in the thinned image which appear to be minutiae which are actually not and these false minutiae have to be removed for the accuracy of further proceedings.

5 Conclusion and Future Scope

In the existing system, once the minutiae are extracted, they are sent for further proceedings like authentication or matching. This reduces the accuracy and efficiency of the system. But, in our system, false minutiae are removed and then it is processed for further proceedings. This increases the perfection of the system with accuracy.

Based on this system, authentication system could be developed in future with a better performance rate.

References

- [1] Tico, M., Kuosmanen, P.: An algorithm for fingerprint image postprocessing. In: Proceedings of the Thirty-Fourth Asilomar Conference on Signals, Systems and Computers, vol. 2, pp. 1735–1739 (November 2000)
- [2] Lee, H.C., Gaensslen, R.E. (eds.): *Advances in Fingerprint Technology*. Elsevier, New York (1991)
- [3] Hrechak, K., McHugh, J.A.: Automated fingerprint recognition using structural matching. *Pattern Recognition* 23, 893–904 (1990)
- [4] Jain, A.K., Bolle, R., Pankanti, S. (eds.): *Biometrics: Personal Identification in Networked Society*. Kluwer Academic Publishers, Boston (1999)
- [5] Jain, L.C., Halici, U., Hayashi, I., Lee, S.B., Tsutsui, S. (eds.): *Intelligent Biometric Techniques in Fingerprint and Face Recognition*. CRC Press, Boca Raton (1999)
- [6] Hastings, E.: A Survey of Thinning Methodologies. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 4(9), 869–885 (1992)
- [7] FuzzyTECH Home Page (September 2003), <http://www.fuzzytech.com>
- [8] Tamura, H.: A Comparison of Line Thinning Algorithms from Digital Geometry viewpoint. In: Proc. of the 4th Int. Conf. on Pattern Recognition, pp. 715–719 (1978)
- [9] James Stephen, M., Prasad Reddy, P.V.G.D.: Implementation of Easy Fingerprint Image Authentication with Traditional Euclidean and Singular Value Decomposition Algorithms. *Int. J. Advance. Soft Comput. Appl.* 3(2) (July 2011) ISSN: 2074-8523

Particle Swarm Optimization Algorithm to Find the Location of Facts Controllers for a Transmission Line

S. Harish Kiran¹, C. Subramani², S.S. Dash³,
M. Arunbhaskar², and M. Jagadeeshkumar²

¹ Department of EEE, Velammal Engineering College,
Chennai, India, 600066

harish99kiran@gmail.com

² Department of EEE, SRM University, Kattankulathure,
Kancheepuram, India, 603203

{csmsrm,m.arunbhaskar}@gmail.com,

mpjagadeesh@yahoo.com

³ Department of EEE, SRM University,

Kattankulathure, Kancheepuram, India, 603203

munu_dash_2k@yahoo.com

Abstract. The main purpose of this paper is to find the optimal location of FACTS controllers in a multi machine power system using enhanced genetic algorithm (EGA) and particle swarm optimization (PSO). Using the proposed method, the location of FACTS controller, their type and rated values are optimized simultaneously. Among the various FACTS controllers, Thyristor Controlled Series Compensator (TCSC) and Unified Power Flow Controller (UPFC) are considered. The proposed algorithms are an effective method for finding the optimal choice and location of FACTS controller and also minimizing the overall system cost, which comprises of generation cost and the investment cost of the FACTS controller using PSO and conventional Newton Raphson's power flow method. A MATLAB coding is developed for Enhanced Genetic Algorithm. In order to verify the effectiveness of the proposed method, IEEE 14- bus system is used. The result obtained in both the algorithm's are compared.

Keywords: Optimal Power Flow (OPF), Flexible AC Transmission System (FACTS), Particle swarm optimization(PSO), Newton Raphson's (NR) power flow.

1 Introduction

In present days with the improvement of the deregulation of electricity market, the traditional practices of power flow in the power system has also been completely changed. To have a better power flow and to have a maximum utilization of the existing power system resources and to increase the power flow in the power system. This can be done by installing a FACTS controller. The placement of the FACTS controller is done on the bases of the economic cost of both the generation and the FACTS controller that is used which becomes essential.

The various parameters considered that are to be controlled by the FACTS controller are transmission line impedances, terminal voltages and angle of the terminal voltage can be controlled by FACTS controllers in an more efficient way. The improvements that have been made in the power system network when the FACTS controllers are included are improvement of steady state in the power flow, system dynamic behavior and enhancement of system reliability. However the other factors that are to be included in the selection of the FACTS controller in the power system are its voltage limits, thermal limits, loop flow, short circuit level and main factor is the sub-synchronous resonance.

The objective of this work is to compare the two algorithms and find the real power allocation of generators and to choose the type and find the optimal and best location for the FACTS controllers such that overall system cost which includes the generation cost of power plants and investment cost of FACTS are minimized. The algorithm used are enhanced Genetic Algorithm (EGA) and particle swarm optimization (PSO) to find the type of the controller to be connected and conventional NR power flow analysis to find the optimal location of the devices and its rating.

2 Controller Selection

The selection of the controller is based on the dynamic and steady state stability of the system. The various problems in the dynamic stability are transient stability, dampening, post contingency and voltage stability and the problems in the steady state stability are voltage limits, thermal limits loop flow, short circuit level and sub-synchronous resonance. For these problems the devices that can be used in all the above mentioned problems are the Thyristor Controlled Series Compensator (TCSC) and Unified Power Flow Controller (UPFC). Another advantage of these controllers is that they have the ability to inject or to observes reactive power and also enhance the power factor in the high voltage transmission line.

3 Cost Functions

As the objective of this paper is to find simultaneously the optimal generation and optimal choice and location of FACTS controllers so as to minimize the overall cost function, which comprises of generation cost and investment costs of FACTS controllers.

3.1 Generation Cost Function

The generation cost function is represented by a quadratic polynomial as follows:

$$C_2(PG) = \alpha_0 + \alpha_1 PG + \alpha_2 PG^2 \quad (1)$$

Where PG is the output of the generator (MW), and α_0 , α_1 and α_2 are cost coefficients.

3.2 Facts Controller Cost Function

Based the Siemens AG Database the cost function for the controller that has been selected to use are as follows:

The cost function for UPFC is:

$$C_{1UPFC} = 0.0003s^2 - 0.2691s + 188.22 \text{ (US\$ / kvar)} \quad (2)$$

The cost function for TCSC is:

$$C_{1TCSC} = 0.0015s^2 - 0.7130s + 153.75 \text{ (US\$ / kVar)} \quad (3)$$

The rating of the device is given by

$$R_{TCSC} = rf * 0.45 - 0.25 \text{ (Mvar)} \quad (4)$$

$$R_{UPFC} = rf * 180 \text{ (MVar)} \quad (5)$$

Where C_{1UPFC} and C_{1TCSC} are in US\$ / kVar and s is the operating of the FACTS controller in MVar. rf is the rated value of each devices. s is the operating range of the FACTS controllers in kVar.

4 Enhanced Genetic Algorithm

In the EGA, the application of the basic genetic operators (parent selection, crossover, and mutation) the advanced and problem-specific operators are applied to produce the new generation. All chromosomes in the initial population are created at random (every bit in the chromosome has equal probability of being switched ON or OFF).

Due to the decoding process selection, the corresponding control variables of the initial population satisfy their upper-lower bound or discrete value constraints. Population statistics are then used to adaptively change the crossover and mutation probabilities. If premature convergence is detected the mutation probability is increased and the crossover probability is decreased. The contrary happens in the case of high population diversity.

4.1 Fitness Function

GAs is usually designed so as to maximize the FF, which is a measure of the quality of each candidate solution. The objective of the OPF problem is to minimize the total operating cost.

Therefore, a transformation is needed to convert the cost objective of the OPF problem to an appropriate FF to be maximized by the GA. The OPF functional operating constraints are included in the GA solution by augmenting the GA FF by appropriate penalty terms for each violated functional constraint. Constraints on the control variables are automatically satisfied by the selected GA encoding/decoding scheme.

Therefore, the GA FF is formed as follows:

$$FF = \frac{A}{\sum_{i=1}^{N_G} Fi(PGi) + \sum_{i=1}^{N_C} \omega_j \cdot Pen_j} \quad (6)$$

$$Pen_j = |h_j(x, u)| \cdot H(h_j(x, u)) \quad (7)$$

Where

- FF fitness function;
- A constant;
- Fi(PGi) fuel cost of unit i
- H(.) Heaviside (step) function;
- NG number of units;
- Nc number of functional operating constraints.

4.2 Enhanced Genetic Algorithm

Step1

Input of the data: vlb, vub, PC, Pm, the function of adaptation and size of the population.

Where:

Vlb, Vub : lower and upper bounds.

PC, Pm : crossover & mutation probabilities

Step2

-To choose arbitrary the initial population.

-To decode the chains to calculate the value of the function to be optimized. For that, it is enough to inject the values of chains decoded in the function.

Step 3

To use the three following operators:

Reproduction.

Crossover.

Mutation.

Step 4

Advanced and problem specific operators

-Hill Climbing

-Gene swap operator

-Gene cross-swap operator

-Gene copy operator

-Gene Inverse Operator

-Gene max-min operator

Step 4

If the convergence of GAs is reached we print the optimal values and stop; else go to the second step

5 Particle Swarm Optimization Algorithm

PSO is a population-based stochastic optimization technique developed by Dr. Eberhart and Dr. Kennedy in 1995, inspired by the social behavior of bird flocking or fish schooling. These phenomena can also be observed on insect colonies, e.g. bees. It is applicable to solving a number of problems where local methods fail or their usage is ineffective, as in this case. One of the most important features of PSO is the ability of optimizing large complex multi-criterial combinatorial problems where the problem with the design of criterial function occurs, for example, it is hard to derive or is not continuous.

PSO however does not need this as it only requires the evaluation of each solution by the fitness function depending on the set of optimized parameters. This function is also used by GA and so is the idea of the initialization of parameter setup as a random generation. The main advantage of PSO compared to GA is the simpler method of providing new solutions based only on two variables - velocity and position related by two linear equations. Each possible solution, represented by a particle flies through the searched space, which is limited by restrictive maximum and minimum values, toward the current optimal position. The particle has its direction and speed of movement (velocity) but it can also randomly decide to move to the best position of all positions or to its own best position. Each particle holds information about its own position (which represents one potential solution), the velocity and the position with the best fitness function it ever has flown through.

5.1 Implementation

The program was implemented in the Matlab environment. The position here represents one potential solution, the velocity shows the trend of this particle, and both parameters are represented by a vector in the program implementation. The particles were coded by natural numbers. The position of each element in the vector space represents the number of the node in which a shunt capacitor should be placed whose value designates the capacitor type. The whole set of particles at a time is called the population. The subset made of newly born particles is called the generation.

The first generation of particles is produced with random position and velocity. Particle velocity is checked whether it is within the limits. The top speed can be different for each unit of velocity vector. If the velocity component exceeds the maximum allowed value, then it is set to the top value. After this correction, the solution is evaluated by the fitness function. The fitness function plays a key role in the program; therefore it is necessary to describe it in more details.

5.2 Fitness and Penalization Functions

The fitness function evaluates the quality of solutions and it incorporates numerous parameters, such as the capital cost of capacitors, expenses covering the power losses in the network per year, and function γ . The power losses are calculated by steady state analysis of the network. The output of the fitness function is total yearly operational costs of the network. The lower the fitness function value, the better the solution. The fitness function is calculated by the following equation:

$$FF = \frac{A}{\sum_{i=1}^{NG} Fi(PGi) + \sum_{i=1}^{Nc} \omega_j \cdot Pen_j} \quad (8)$$

Where

- FF fitness function;
- A constant;
- Fi(PGi) fuel cost of unit i
- H(.) Heaviside (step) function;
- NG number of units;
- Nc number of functional operating constraints.

$$Pen_j = |h_j(x, u)| \cdot H(h_j(x, u)) \quad (9)$$

5.3 PSO Algorithm Flow Sequence

Algorithm PSO

Begin

- Generate random population of N solutions(particles);
- For each individual $I \in N$ calculate fitness (i);
- Initialize the value of the weight factor ω ;
- For each particle;
 - Set pBest as the best position of particle i;
 - If fitness (i) is better than pBest;
 - pBest(i)=fitness (i);
- End;
- Set gBest as the best fitness of all particles;
- For each particle;
 - Calculate particle velocity according to Eq. (6a);
 - Update particle position according to Eq. (6b);
- End;
- Update the value of the weight factor ω (option);
- Check if termination=true;

End

6 Test Results

A MATLAB coding is developed for particle swarm optimization. In order to verify the effectiveness of the proposed method IEEE 14 bus system is used. Different operating conditions are considered for finding the optimal choice and location of FACTS controllers.

The total population size is selected as 150, the mutation probability as 0.01 and crossover probability as 1.0. The following tabulation table 3 is the result obtained when the PSO coding was tested for the IEEE 14 bus system.

From the tabulation it is found that the bus number 4 and 6 are repeated twice and more times. After careful study from the tabulation it is said to connect UPFC on the bus number 6 connecting to line number 10 and line number 11 with a rating of 0.2pu to 0.5pu.

Table 1. PSO Simulation Result Obtained

S.I No.	Pg1	Pg2	Pg3	Loss	Optimal		
					Device	Location (Ns –Nr)	Rating
1	141.88	75.96	55.35	14.20	TCSC	11(6-11)	0.3713
2	153.69	63.50	56.27	14.42	UPFC	9(4-9)	-0.888
3	146.98	70.42	55.95	14.35	UPFC	10(5-6)	0.3279
4	151.75	65.97	55.56	14.30	UPFC	9(4-9)	-0.749

As the EGA has the ability to multiple combination of the selection of the controller each time it is simulated, hence it does not have a constant selection of the controller there will be a continuous change in the location, rating of the controller and the line to which it has to be connected in the power system network. The following table 4 shows the result which were obtained when the Enhanced Genetic coding where run.

Table 2. Shows the Results of Enhanced Genetic Algorithm Simulation Obtained

S.I No.	Pg1	Pg2	Pg3	Loss	Optimal		
					Device	Location (Ns –Nr)	Rating
1	157.33	55.03	63.35	16.73	TCSC	9(4-9)	-0.977
2	118.62	82.39	77.89	19.91	UPFC	17(9-14)	-0.682
3	157.33	55.04	63.35	16.73	TCSC	9(4-9)	-0.976
4	95.94	141.63	31.78	10.37	UPFC	5(2-5)	-0.352

From the above tabulation it is found that the number of times the bus number 9 and bus number 4 repeated is four times. Hence it is desired to connect TCSC on line number 9 connecting the bus number 4 to bus number 9 with a rating range of 1pu to -1pu.

Hence from the table 1 and table 2 it is evident that the location of the device, type of the devices and the rating of the devices keeps on changing continuously each time the load flow program is simulated.

7 Conclusions

Based on the results obtained by the simulation of Enhanced Genetic Algorithm it's found that the use of TCSC with a rating of 1.0pu to -1.0pu at the line number 9 connecting bus number 4 to bus number 9 will give an optimum power flow solution. As, in the case of Particle Swarm Optimization algorithm it's found to use UPFC with a rating of 0.2pu to 0.5pu on bus number 6 connecting between the line number 10 and line number 11 which will give an optimum power flow solution.

Based on the cost effect it is ideal to use UPFC of the operating range -1pu to 1pu at bus number 6 connecting the line number 10 and line number 11.

References

1. Abdelsalam, H.A., Aly, G., Abdelkrim, M., Shebl, K.M.: Optimal location of the unified power flow controller in electrical power systems. In: Proc. IEEE Power Eng. Soc. Power Systems Conf. Expo., vol. 3, pp. 1391–1396 (October 2004)
2. Fang, W.L., Ngan, H.W.: Optimizing location of unified power flow controllers using the method of augmented Lagrange multipliers. In: Proc. Inst. Elect. Eng., Gen., Transm., Distrib., vol. 146(5), pp. 428–434 (September 1999)
3. Glover, J.D., Sarma, M.: Power System Analysis & Design, 3rd edn. PWS-Kent, Boston (2002)
4. Habur, K., O’Leary, D.: For cost effective and relative transmission of electrical energy (reviewed by Messrs)
5. Syafrullah, M., Salim, N.: Improving Term Extracting Particle Swarm optimization Techniques. Journal of Computer Science 6(3), 323–329 (2010)
6. Hingorani, N.G.: Understanding FACTS -Concepts and Technology of Flexible AC Transmission Systems. Standard Publishers Distributors, Delhi- 110 006
7. Noroozian, M., Angquist, L., Ghandhari, M., Andersson, G.: Use of UPFC for optimal power flow control. IEEE Trans. Power Del. 12(4), 1629–1634 (1997)
8. Paar, M., Toman, P.: Utilization of particle Swarm Optimization Algorithm for Optimization of MV Network Compensation. This Work was Supported in Part Ministry of Education, Youth and Sports of the Czech Republic Under Project No. MSM0021630516 (April 2007)
9. An, S., Gedra, T.W.: UPFC ideal transformer model. In: Proc. North Amer. Power Symp., pp. 46–50 (October 2003)
10. Vijayakumar, K., Kumudinidevi, R.P.: A new method for optimal location of FACTS controllers using genetic algorithm. Journal of Theoretical and Applied Information Technology (January 2007)

Uncertain Data Classification Using Rough Set Theory

G. Vijay Suresh¹, E. Venkateswara Reddy¹, and E. Srinivasa Reddy²

¹ Universal College of Engineering & Technology, Guntur
vijaysuresh.g@gmail.com,
evr_universal@yahoo.com

² University College of Engineering & Technology,
ANU, Guntur
edra_67@yahoo.com

Abstract. Data uncertainty is common in real-world applications due to various causes, including imprecise measurement, network latency, out-dated sources and sampling errors. As a result there is a need for tools and techniques for mining and managing uncertain data. In this paper proposes a Rough Set method for handling data uncertainty. Rough set is a mathematical theory for dealing with uncertainty. Uncertainty implies inconsistencies, which are taken into account, so that the produced are categorized into certain and possible with the help of rough set theory. Experimental results show that proposed model exhibits reasonable accuracy performance in classification on uncertain data.

1 Introduction

Data mining and knowledge discovery techniques are widely used in various applications in business, government, and science. Examples include banking, bioinformatics, environmental modeling, epidemiology, finance, marketing, medical diagnosis, and meteorological data analysis [1] [2] [3]. Data is often associated with uncertainty because of measurement inaccuracy, sampling discrepancy, outdated data sources, or other errors. [1][2] Uncertainty can be caused by our limited perception or understanding of reality (e.g., limitations of the observation equipment; limited resources to collect, store, transform, analyze, or understand data). It can also be inherent in nature (e.g., due to prejudice). Moreover, sensors (e.g., acoustic, chemical, electromagnetic, mechanical, optical radiation and thermal sensors) are often used to collect data in applications such as environment surveillance, security, and manufacturing systems. Data uncertainty can be categorized into two types, namely existential uncertainty and. value uncertainty [2] [3]. In the first type it is uncertain whether the object or data tuple exists or not. For example, a tuple in a relational database could be associated with a probability value that indicates the confidence of its presence. In value uncertainty, a data item is modelled as a closed region which bounds its possible values, together with a probability density function of its value. All these scenarios lead to huge amounts of uncertain data in various real-life situations.

2 Research Background

2.1 Rough Set Concepts

Rough set theory can be viewed as a specific implementation of Frege’s idea of uncertainty [8] i.e., imprecision in this approach is expressed by a boundary region of a set, and not by a partial membership, like in fuzzy set theory. Rough set concept can be defined quite generally by means of topological operations, *interior* and *closure*, called *approximations*. Let us describe this problem more precisely. Suppose we are given a set of objects U called the *universe* and an indiscernibility relation $R \subseteq U \times U$, representing our lack of knowledge about elements of U . For the sake of simplicity we assume that R is an equivalence relation. Let X be a subset of U . We want to characterize the set X with respect to R . To this end we will need the basic concepts of rough set theory given below [9][10].

- The *lower approximation* of a set X with respect to R is the set of all objects, which can be for *certain* classified as X with respect to R (are *certainly* X with respect to R).
- The *upper approximation* of a set X with respect to R is the set of all objects which can be *possibly* classified as X with respect to R (are *possibly* X in view of R).
- The *boundary region* of a set X with respect to R is the set of all objects, which can be classified neither as X nor as not- X with respect to R .

Now we are ready to give the definition of rough sets.

- Set X is *crisp* (exact with respect to R), if the boundary region of X is empty.
- Set X is *rough* (inexact with respect to R), if the boundary region of X is nonempty.
- Formal definitions of approximations and the boundary region are as follows:
R-lower approximation of X

$$\underline{R}(x) = \bigcup_{x \in U} \{R(x) : R(x) \subseteq X\} \tag{1}$$

- *R-upper approximation of X*

$$\overline{R}(x) = \bigcup_{x \in U} \{R(x) : R(x) \cap X \neq \emptyset\} \tag{2}$$

- *R-boundary region of X*

$$RN_R(X) = \overline{R}(X) - \underline{R}(X) \tag{3}$$

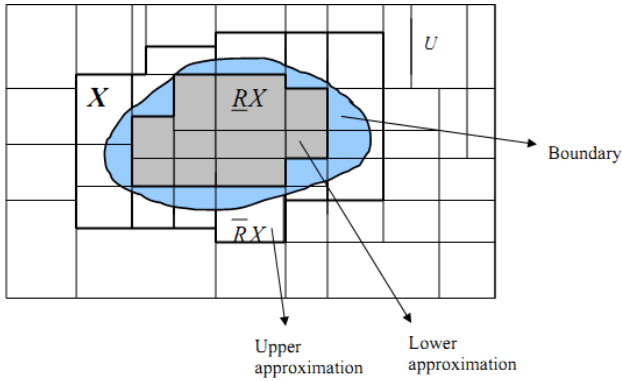


Fig. 1. Representation of the data partitioning for a subset X

2.2 Indiscernibility and Approximation

An information system is a data table containing rows labeled by objects of interest, columns labeled by attributes and entries of the table are attribute values. For example, a data table can describe a set of patients in a hospital. The patients can be characterized by some attributes, like *age*, *sex*, *blood pressure*, *body temperature*, etc. With every attribute a set of its values is associated, e.g., values of the attribute *age* can be *young*, *middle*, and *old*. Attribute values can be also numerical. In data analysis the basic problem we are interested in is to find patterns in data, i.e., to find a relationship between some set of attributes, e.g., we might be interested whether *blood pressure* depends on *age* and *sex*.

Central to RST is the concept of indiscernibility. Let $I = (U, A)$ be an information system, where U is a non-empty set of finite objects (the universe of discourse) and A is a non-empty finite set of attributes such that $a : U \rightarrow V_a$ for every $a \in A$. V_a is the set of values that attribute a may take. For any $P \subseteq A$, there is an associated equivalence relation $IND(P)$.

$$IND(P) = \{(x, y) \in U^2 \mid \forall a \in P, a(x) = a(y)\} \tag{4}$$

The partition of U , generated by $IND(P)$, is denoted by $U/IND(P)$ and can be defined as follows:

$$U/IND(P) = \otimes \{a \in P : U/IND(\{a\})\} \tag{5}$$

where

$$U/IND(\{a\}) = \{\{x \mid a(x) = b, x \in U\} \mid b \in V_a\} \tag{6}$$

and

$$A \otimes B = \{X \cap Y : \forall X \in A, \forall Y \in B, X \cap Y \neq \emptyset\} \tag{7}$$

If $(x, y) \in IND(P)$, then x and y are indiscernible by attributes from P . The equivalence classes of the P -indiscernibility relation are denoted by $[x]_P$ [15]

Table 1. Information Table

ID	Temp	BP	Class HP
A	N	L	N
B	N	L	N
C	N	M	Y
D	H	M	N
E	H	H	Y
F	H	H	Y
G	H	H	N

$$Temp \in \left\{ \begin{array}{l} Normal(N) \\ High(H) \end{array} \right\}$$

$$BP \in \left\{ \begin{array}{l} Low(L) \\ Medium(M) \\ High(H) \end{array} \right\}$$

$$HP(Class) \in \left\{ \begin{array}{l} No(N) \\ Yes(Y) \end{array} \right\}$$

Equivalence relations induce a partition of the elements, i.e. a set of equivalence classes. For example, consider the relation **R** on the objects (identified by their ids in the table), defined as follows: (x,y) in **R** if and only if x and y have the same value for Temp. This induces the partition: $\{ABC\},\{DEFG\}$, since ABC all have value N and $DEFG$ all have value H for temperature. We denote the partition of the objects based on temperature or *induced by Temp*, using the notation:

$$R_{Temp} = \left\{ \begin{array}{l} \{ABC\} \\ N \end{array} \right\} \left\{ \begin{array}{l} \{DEFG\} \\ H \end{array} \right\}$$

Similarly the partition induced by BP is given as follows:

$$R_{BP} = \left\{ \begin{array}{l} \{AB\} \\ L \end{array} \right\} \left\{ \begin{array}{l} \{CD\} \\ M \end{array} \right\} \left\{ \begin{array}{l} \{EFG\} \\ H \end{array} \right\}$$

It is possible to create a partition induced by more than one attribute. For example the partition induced by both Temp and BP (i.e., the relation **R** is that objects x and y have the same class if they have the same values for both Temp and BP), is given as follows

$$R_{Temp} \cap BP = \left\{ \left\{ \begin{array}{l} \{AB\} \\ NL \end{array} \right\} \left\{ \begin{array}{l} \{C\} \\ NM \end{array} \right\} \left\{ \begin{array}{l} \{D\} \\ HM \end{array} \right\} \left\{ \begin{array}{l} \{EFG\} \\ HH \end{array} \right\} \right\}$$

We then label each partition by the values from the two attributes, for instance $\{AB\}$ with label NL means, both A and B , have value N for $Temp$ and L for BP .

3 Related Work

Let us consider the data table which is uncertain in nature:

Table 2. Information Table

ID	BP	Temp	Class HP
A	L	?	N
B	?	H	N
C	H	H	Y
D	H	?	Y

The Table.2 consists of some missing values this can be avoided by using different approaches

- Ignoring examples with unknown values of attributes [16],
- Assuming additional special value for an unknown value of attributes,
- *Using probability theory.* For example, using relative frequencies of known values of a given attribute A for assigning them to unknown values [17].

3.1 For Missing Value Classification

The main idea to use of rough set theory is to generate the certain and uncertain rules. All results of uncertainty are manifested finally by inconsistent information in the decision table. The main idea of the method is to replace each example with an unknown value of attribute A by the set of examples, in which attribute A has its every possible value. Thus, if attribute A has an unknown value for example E , and attribute A has m possible values, then E will be replaced by m new examples $E', E'', \dots, E^{(m)}$. [19]. When example E has two unknown values of attributes A and B , and there is m possible values of A and n possible values of B , then E will be replaced by $m \cdot n$ examples, and so on. The most obvious rationale of the method is the following: since the value of an attribute A for a given example E is unknown, every possible value of A is considered, and every such value corresponds to a new example. On the other hand, the fact that attribute A has an unknown value for example E , and that E is a member of some class C may be interpreted in yet another way: an expert classified E as a member of class C not knowing the value of A , i.e., that such a value was not necessary for classification. This implies that it does not matter what a value it was, hence, A may assume any value from its domain.

Let us fill the missing values in Table.2 with all given possibilities.

Table 3. Information Table

ID	Temp	BP	Class HP
A'	L	M	N
A''	L	H	N
B'	L	H	N
B''	H	H	N
B'''	M	H	N
C	H	H	Y
D'	H	M	Y
D''	H	H	Y

The pairs of examples (B'', C) and $((B'', D''))$ are inconsistent. To avoid this we implement the Rough set to resolve these inconsistencies. From Table.4, the partition P^* is equal to $\{\{A'\}, \{A'', B'\}, \{B'', C, D''\}, \{B'''\}, \{D'\}\}$, where $P = \{BP, Temp\}$. The lower approximation of class $\{A', A'', B', B'', B'''\}$, corresponding to N value of Class HP is $\{A', A'', B', B'''\}$. Similarly, the lower approximation of the class $\{C, D', D''\}$ is $\{D'\}$. The upper approximation of the class $\{A', A'', B', B'', B'''\}$ is

{A', A'', B', B'', B''', C, D''}, and the upper approximation of the class {C, D', D''} is {B'', C, ', D''}. Thus, for both classes, error ϵ is the same and equal to

$$\frac{7-4}{8} = \frac{4-1}{8} = 0.375$$

The expression for an error may be interpreted as follows: three examples (*B''*, *C* and *D''*) out of eight are possibly but not certainly correctly classified. The error would be 0.375 when all three examples: B'', C, and D'' are mistakenly classified, i.e., when none of them belong to the corresponding class. To avoid this we will consider two tables *Table.4* and *Table.5*

Table 4. Information Table

ID	Temp	BP	Class HP
A'	L	M	N
A''	L	H	N
B'	L	H	N
B''	H	H	Y
B'''	M	H	N
C	H	H	Y
D'	H	M	Y
D''	H	H	Y

Table 5. Information Table

ID	Temp	BP	Class HP
A'	L	M	N
A''	L	H	N
B'	L	H	N
B''	H	H	N
B'''	M	H	N
C	H	H	N
D'	H	M	Y
D''	H	H	N

Note that *Tables4&5* are consistent. Thus, from *Table 4*, rules describing the class {A', A'', B', B'''}, i.e. certain rules for *NO* value of *Class HP*, may be induced as follows:

$$\begin{aligned} (BP, Low) &\rightarrow (Class\ HP, NO), \\ (BP, Medium) &\rightarrow (Class\ HP, NO). \end{aligned}$$

Similarly, from *Table 5*, the rule for the class {D'}, i.e. a certain rule for *Yes* value of *Class HP*, is induced:

$$(BP, High) \wedge (Temp, Normal) \rightarrow (Class HP, YES).$$

Upper approximations of the classes imply Tables 4 and 5 (these tables are the same as implied by lower approximations because Class HP has two values). From Table 5, rules for the class $\{A', A'', B', B'', B''', C, D''\}$, i.e., possible rules for NO value of Class HP induced:

$$\begin{aligned} (BP, Low) &\rightarrow (Class HP, NO), \\ (Temp, High) &\rightarrow (Class HP, NO), \\ (BP, Medium) &\rightarrow (Class HP, NO). \end{aligned}$$

Finally, from Table 5, a rule for the class $\{B'', C, D', D''\}$, i.e., a possible rule for YES value of Class HP is induced:

$$(BP, High) \rightarrow (Class HP, Yes).$$

The certain rules, listed above, are absolutely correct—no error analysis is required. The error for the possible rules is always smaller than 37.5%. The above rules, certain and possible, are presented in the minimal discriminate form [18].

4 Experiment and Results

4.1 Data

The expression patterns of 27 markers were assessed in a series of 261 adenocarcinomas. 12 markers were scored as either present or absent (+ or -). The remaining markers showed variation in intensity between tumors and were scored as weak, intermediate or strong (0, 1, 2 or 3). Furthermore, Undefined indicates cores which are missing and therefore cannot be scored. To predict the site of origin using the expression profile of the 27 candidate markers taken from the secondary tumor. The ROSETTA[20],[21] system is a software package for inducing rough-set based rule. In addition to the core features described there, the system includes a large number of algorithms for discretization, reduct computation, and rule pruning and classifier evaluation.

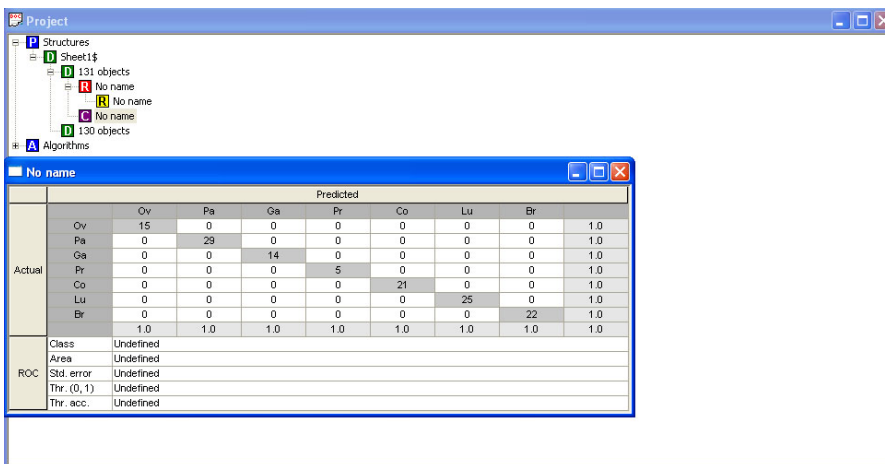


Fig. 2. Classification Results

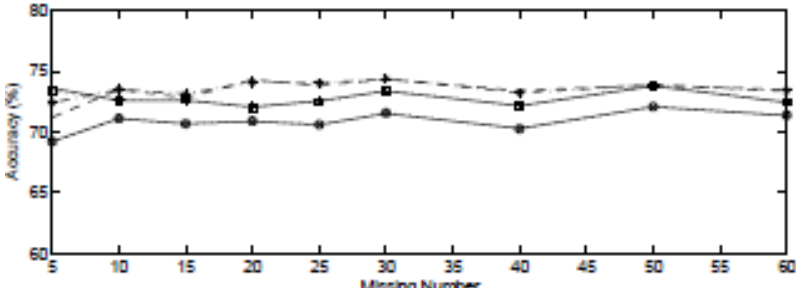


Fig. 3. Comparison Figure data adenocarcinomas

The confusion matrix shows the overall accuracy (i.e. 0.715385), as well as the sensitivity and accuracy for each class. For example, the *Ov* decision class has a sensitivity of 0.54 (i.e. of $7+5+1 = 13$ objects actually belonging to *Ov*, 7 was correctly classified as *Ov*: $7/13 = 0.54$) and an accuracy of 0.88 (i.e. of $7+1 = 8$ objects predicted to *Ov*, 7 were actually belonging to this class: $7/8 = 0.88$).

5 Conclusion

In this paper, we propose very simple method to deal with unknown values of attributes: every example with unknown values of attribute *A* is replaced by the set of examples having every possible value for *A*. This is the most conservative approach because an unknown value is replaced by every possible value. This method produces, in general, inconsistent decision tables. However, the problem of learning rules from inconsistent examples may be easily solved using rough set theory. Thus, two different sets of rules are computed: certain and possible. Certain and possible rules may be propagated separately during an inference process in an expert system, producing thus new certain and possible rules, respectively. Therefore, the inference engine of an expert system may be divided into two parallel subsystems, for certain and possible rules, in which certain and possible rules are processed separately. We plan to explore more classification approaches for various uncertainty models and find more efficient training algorithms in the future.

References

1. Leung, C.K.-S.: Mining uncertain data. In: WIREs Data Mining and Knowledge Discovery, vol. 1, p. 2. John Wiley & Sons, Inc. (2011)
2. Suresh, G.V., Shaik, S., Reddy, E.V., Shaik, U.A.: Gaussian Process Model for Uncertain Data Classification. International Journal of Computer Science and Information Security (IJCSIS) 8(9), 111–115 (2010)
3. Chau, M., Cheng, R., Kao, B.: Uncertain Data Mining: A New Research Direction. In: Proceedings of the Workshop on the Sciences of the Artificial, Hualien, Taiwan, December 7-8 (2005)

4. Aggarwal, C.C.: *Managing and Mining Uncertain Data*. Kluwer Academic Publishers, Boston
5. Aggarwal, C.C.: A Survey of Uncertain Data Algorithms and Applications. *IEEE Transactions on Knowledge and Data Engineering* 21(5) (2009)
6. Aggarwal, C.C., Yu, P.S.: Outlier detection with uncertain data. In: *SDM*, pp. 483–493. SIAM (2008)
7. Hamdan, H., Govaert, G.: Mixture Model Clustering of Uncertain Data. In: *IEEE International Conference on Fuzzy Systems*, pp. 879–884 (2005)
8. Pawlak, Z.: Rough sets. *Int. J. of Information and Computer Sciences* 11(5), 341–356 (1982)
9. Pawlak, Z., Skowron, A.: Rough membership function. In: Yeager, R.E., Fedrizzi, M., Kacprzyk, J. (eds.) *Advances in the Dempster-Schafer of Evidence*, pp. 251–271. Wiley, New York (1994)
10. Voges, K.E.: *Research Techniques Derived From Rough Sets Theory: Rough Classification and Rough Clustering* (2005)
11. Olve Maudal, Y.: Preprocessing data for Neural Network based Classifiers: Rough Sets vs Principal Component Analysis. Project report, Department of Artificial Intelligence, University of Edinburgh (1996) 16, 20, 41, 42, 47, 60
12. Pawlak, Z.: Rough sets. *International Journal of Computer and Information Sciences* 11(5), 341–356 (1982) ISSN 0091-7036. 2
13. Pawlak, Z.: *Rough Sets: Theoretical Aspects of Reasoning About Data*. Kluwer Academic Publishers, Dordrecht (1991) 2, 15, 16, 69, 70
14. Mac Parthalain, M., Shen, Q.: On rough sets, their recent extensions and applications. *The Knowledge Engineering Review* 25(4), 365–395 (2010)
15. Quinlan, J.R.: Induction of decision trees. *Machine Learning* 1, 81–106 (1986)
16. Knonenko, I., Bratko, I., Roskar, E.: Experiments in automatic learning of medical diagnostic rules. Technical Report, Jozef Stefan Institute, Ljubljana, Yugoslavia (1984)
17. Michalski, R.S.: A theory and methodology of inductive learning. In: Michalski, R.S., Carbonell, J.G., Mitchell, T.M. (eds.) *Machine Learning. An Artificial Intelligence Approach*, pp. 83–134. Morgan Kaufmann (1983)
18. Grzymala-Busse, J.W.: On the Unknown Attribute Values in Learning from Examples. In: *Proc. of the ISMIS 1991, 6th International Symposium on Methodologies for Intelligent Systems*, Charlotte, North Carolina, October 16–19, pp. 368–377 (1991)
19. Komorowski, J., Øhrn, A., et al.: The ROSETTA Rough Set Software System. In: Klösgen, W., Zytkow, J. (eds.) *Handbook of Data Mining and Knowledge Discovery*, pp. I554–I559. Oxford University Press (2002)
20. Andersson, R., Vesterlund, J.: *GENOMIC ROSETTA - Application Mode User Manual*. The Linnaeus Centre for Bioinformatics, Uppsala (2005)
21. Hastie, T., Tibshirani, R.J., et al.: *The Elements of Statistical Learning*. Springer, New York (2001)
22. Quinlan, J.R.: Probabilistic decision trees. In: Kodratoff, Y., Michalski, R.S. (eds.) *Machine Learning. An Artificial Intelligence Approach*, vol. III, pp. 140–152 (1990)
23. Yasdi, R., Ziarko, W.: An expert system for conceptual schema design: A machine learning approach. *Int. J. Man-Machine Studies* 29, 351–376 (1988)
24. Jacobs, I.S., Yao, Y.Y.: A step towards the foundations of data mining. *Data Mining and Knowledge Discovery: Theory, Tools, and Technology V*. The International Society for Optical Engineering, 254–263 (2003)

Design of Composite Web Service to Obtain Best QoS

Urjita Thakar and Abhishek Agrawal

Department of Computer Engineering, Shri G.S. Institute of Technology & Science,
23-Park Road, Indore-452003 (MP), India

urjita@rediffmail.com, abhishek_abhiitm@yahoo.com

Abstract. In the present Web services scenario, user demands are complex in nature and cannot be answered by a single Web Service. Composite services need to be constructed to fulfill such demand. QoS is an important aspect of service composition. The overall QoS of the composite service may be decided by the QoS offered by the constituting component services. Therefore a complex service should be formed using component services with matching QoS values. In this paper, a method has been proposed to form complex service sets using component services with matching QoS values such that user's cost requirement is also satisfied. Availability, response time and throughput are important QoS parameters. Very few number of sets are presented to the user in the form of a list such that the set with best QoS appears at the top. The proposed method thus is very useful for the user and enables him to avail the complex service with best QoS.

Keywords: Web Services, Composition, QoS (Quality of Service).

1 Introduction

The future perspective of the Internet is being driven by a new concept commonly known as Web Services [1]. Web services are applications that can be published, located, and invoked across the Internet. These are based on Service Oriented Architecture [2].

At present, large number of Web Services are present on the World Wide Web. Most of these are designed to serve a specific type of business functionality. The present needs of business enterprises are very huge in nature and can't be served by a single web service. Therefore, composition of several web services to form a complex Web service is required.

For a complex Web service to perform well, the component services constituting them should match well with each other. Different services are offered with different values for quality of service parameters by different providers. Some providers offer high value for a particular QoS parameter, while some other providers may offer low value for the same parameter. The overall QoS of the composite service may be decided by the worst QoS value of a constituting component. Availability, response time and throughput are important QoS parameters that indicate performance of a service and are therefore also important for the service requester. In this paper, an approach has been proposed for determining the components that shall constitute a composite service with best QoS.

Rest of the paper is organized as follows: The related work is discussed in section 2. The proposed method is presented in section 3. In section 4, testing and results of the presented method have been discussed. The paper is concluded in section 5.

2 Related Work

In this section, some contributions by various researchers relevant to this paper are discussed. A solution for dynamic web service composition has been discussed by Ming et al. in their work [3]. The method corresponds to finding matching services from a pool of services. The user's requirement was broken down into a series of abstract web services. By semantic matching among the abstract services, atomic services or the other smaller composite services, a web service composition is obtained for execution.

Senkul et al. have proposed a system that can compose web services under the user's constraints on the overall composite service as well as requirements on the atomic services [4]. On the basis of the constraints and acceptance levels, a set of prioritized feasible plans is generated and ranked. Boumhamdi et al. have proposed architecture for dynamic composition of Web services as per user's requirements and availability of resources [5]. This architecture also has the ability to re-configure the composite service at runtime in case of some failure.

Ye et al. have proposed a QoS Broker for discovering Web Services based on QoS parameters [6]. Al-Masri et al. developed Web Service Relevancy Function (WsRF) for measuring the relevancy ranking of a particular Web service based on client's preferences and QoS metrics [7]. The Ranking function finds the best available Web service during service discovery process based on a set of given client QoS preferences.

3 Proposed Method

In this section, the proposed method for composing web services based on QoS parameters- Availability, Response time and throughput is discussed. In section 3.1, the method for calculating the values for these parameters is presented. The algorithm to design the composite Web Services with best QoS parameters is presented in section 3.2.

3.1 Calculation of Values for QoS Parameters

The QoS requirements for web services mainly refer to the quality aspect of a web service with regard to performance, reliability, scalability, capacity, robustness, accuracy, integrity, accessibility, availability, response time, throughput, interoperability [8][9][10].

In this paper, QoS parameters used for composition of Web Service are availability, response time and throughput. The proposed system calculates the values for these QoS parameters of each service provider as discussed below.

Availability. To find the availability of services, system takes endpoint URL of a service from the service registry and generates a request for each service. If a reply is received from a service in expected time then that service is considered to be available. The number of successful invocation and total invocation for that service are incremented by 1. Otherwise value of total invocation for that service is incremented by 1. This process is repeated periodically after few minutes by the system. Availability of the service is calculated using following equation.

$$\text{Availability} = \text{Number of successful invocation} / \text{Total invocation}$$

Response Time. To find the response time of a service, system notes the time of sending the request to the service. The time of arrival of response is also noted. The response time of the service is calculated as given below.

$$\text{Response Time} = \text{Time of receiving Response} - \text{Time of making the request}$$

Throughput. To find the throughput of services, system generates a number of requests for a particular service for a fixed time period. The corresponding responses from the service are captured. Throughput of the service is calculated using following equation.

$$\text{Throughput} = \text{Total number of handled requests} / \text{time}$$

The proposed algorithm is discussed next.

3.2 Proposed Algorithm

Following algorithm is proposed for composition of Web service to obtain best QoS. In this algorithm, maximum cost value payable for a complex service is used as threshold value to obtain complex service sets. Certain iterations are followed to get complex service sets. In first iteration, two services with different functionalities that appeared in the business process are considered and total cost of this service set is compared with the threshold. If it is less than or equal to the threshold cost then that service set is accepted, otherwise rejected. In the next iterations accepted service sets are combined with other services with different functionality one by one to obtain next service sets. Total cost of these service sets is compared with the threshold cost to determine acceptable sets. These iterations are repeated till complex service sets containing all required component services with different functionalities for the desired business process are obtained.

For all the acceptable service sets, normalized QoS values are determined. The worst value is determined for each QoS parameter for each service set. Next, the average of normalized QoS values is calculated for each service set.

These service sets are arranged in the form of a list such that the set with best QoS appears at the top. It enables the user to select best composite web service with best QoS. The algorithm is as discussed below.

Algorithm

Input: Number of services, Number of Service Providers for each service, Service Providers for each type of service, User's Readiness to pay, Cost of each service provider, QoS parameter values for each Service Provider, Required number of composite Web services.

Output: Composite Web Services (final[])

QoSBasedComposition()

- (1) Declare variables arr1[], arr2[] and arr3[]
//arr1[], arr2 and arr3[] stores service sets
- (2) set:arr1[]=All Service Provider of 1st type of service
- (3) for (i = 2 to Number of Services to be composed)
- (4) set:arr2[]=All Service Provider of ith type of service
- (5) for(j = 1 to number of service sets in arr1[])
- (6) for(k = 1 to number of service providers in arr2[])
- (7) arr3[] = arr1[j] + arr2[k]

```

//Generates new service set by combining
service providers of arr1[j] with service
providers of arr2[k]
(8) End of step 6 loop
(9) End of step 5 loop
//Gets number of service sets. Each service set
is combination of i type of services.
(10) clear arr1[]
//Delete data stored in arr1[]
(11) arr1[] = MaxCost(arr3[])
//Checks service sets of arr3[] satisfying
user's readiness to pay and stores in arr1[]
(12) clear arr3[] and arr2[]
//Delete data stored in arr3[] and arr2[]
(13) End of step 3 loop
//Generates service sets having Service
Providers of all types and satisfying user's
readiness to pay
(15) arr3[] = MinQoS(arr1[])
//Finds QoS value of service sets in arr1[]
(16) for (i = 1 to Required number of composite Web
services)
(17) final[i] = arr3[i]
(18) End of step 16 loop
//Required number of composite Web services
with highest QoS values is resulted.

```

MaxCost(arr3[] ServiceSets)

```

(1) Declare variable arr1[].
//arr1[] contains service sets.
(2) for (i = 1 to Number of service sets in arr3[])
(3) if (total cost of service set arr3[i] <= Readiness to
pay)
(4) arr1[] = arr3[i]
//Service set arr3[i] is satisfying cost
constraints, therefore stored in arr1[].
(5) End of if
(6) End of step 2 loop
(7) return arr1[]

```

MinQoS(arr3[] ServiceSetsSatisfyingUser'sReadinessToPay)

```

(1) Declare variable flag, arr1[], arr2[][] qos[]
//arr1[] stores service sets, arr2[][] stores
component service providers of each service
set, qos[] stores QoS value of service sets
(2) for (i = 1 to Number of service sets in arr3[])
(3) Parse component service providers from service set
arr3[i]
(4) for(j = 1 to Number of Component service providers in
service set arr3[i])

```

```

(5) arr2[i][j] = jth component service provider of service
    set arr3[i]
(6) End of step 4 loop
(7) qos[] = min(QoS parameter values of all component
    services of service set arr2[i][j])
    //Finds minimum QoS parameter value among QoS
    parameter value of all component services of
    service set arr2[i][j]
(8) End of step 2 loop
(9) for (i = 1 to Number of service sets in arr1[])
(10) for (j = i+1 to Number of service sets in arr1[])
(11) if (qos[i] < qos[j])
    //Selection Sort is applied to arrange service
    sets in descending order of their QoS values
(12) Exchange arr1[i] with arr1[j]
(13) Exchange qos[i] with qos[j]
(14) End of if
(15) End of step 10 loop
(16) End of step 9 loop
(17) return arr1[]
    
```

The proposed algorithm has been tested by taking a test case as discussed next.

4 Testing, Results and Discussion

In this paper, for testing purpose a complex service called as “Tour Support System” is considered. This service may be obtained by composition of one or more services for Travel, Hotel and Pickup. The user inputs the services required in the composition, required parameters of those services and his willingness to pay. Let the user’s requirements for Travel, Hotel and Pickup service be as shown in table 1.

Table 1. User’s Requirements

Travel Service	Hotel Service:	Pickup Service:
Source : Bhopal	City : Gwalior	City : Gwalior
Destination : Gwalior	Hotel Type : 2-star	Vehicle Name : Qualis
Travel Mode : Car	AC Required : AC	AC Required : AC
AC Required : AC	Single/Double : Double	No. of Rooms : 1
No. of Seats : 4	No. of Rooms : 2	Date From :15:06:2011
Date of Journey :15:06:2011	Date From :15:06:2011	Date To :17:06:2011
	Date To :17:06:2011	
Ready to pay cost for Composite Web Service : 17700 units		

As per user’s requirements, system finds service providers and their cost from the service registry. Table 2 shows the list of component service providers discovered from the registry, the cost charged by them and QoS values. The system continuously calculates the values for various QoS parameters. Since the QoS parameters have different units and also for some parameters high value is considered to be good while for other, low value is considered to be better, QoS values are normalized as shown in the table given below.

Table 2. Table of the discovered service providers, Cost charged by them, QoS values and normalized QoS values

Service Provider	Cost	Response Time		Availability		Throughput	
		Calculated	Normalized	Calculated	Normalized	Calculated	Normalized
TravelA	3600	180	0.5166	86.66	0.8666	851	1
TravelB	3600	100	0.93	100	1	812	0.9554
TravelC	3200	93	1	100	1	839	0.9858
TravelG	3600	185	0.5027	100	1	837	0.9835
HotelD	10200	189	0.5714	76.66	0.7666	416	1
HotelF	9600	108	1	100	1	405	0.9735
HotelG	10800	167	0.6467	93.33	0.9333	393	0.9447
PickupA	3900	74	1	100	1	813	0.9475
PickupB	3900	183	0.4043	100	1	831	0.9668
PickupD	4200	108	0.6851	83.33	0.8333	857	0.9988
PickupE	4200	165	0.4484	96.33	0.9666	833	0.9708
PickupG	4200	146	0.5068	100	1	733	0.8543
PickupJ	3900	169	0.4378	66.66	0.6666	858	1

In this test case, when the cost of TravelA, HotelD and PickupD are added, the cost is calculated to be 18000 which is greater than 17700 which is user’s readiness to pay. Thus the combination TravelA, HotelD and PickupD is not considered for composition. When TravelC, HotelF and PickupA are taken, then total cost calculated is 16700 which is less than user’s readiness to pay. Thus this combination is considered for complex service set.

For finding QoS parameters of this service set, system finds minimum normalized values for availability, response time and throughput as 1, 1 and 0.9475 respectively. Average of the minimum QoS parameters is 0.9825.

This is repeated for all the service combinations meeting cost requirements. The Table 3 shows 10 service sets sorted such that the set with best QoS appears at the top.

Table 3. Sorted List of Composite Web services based on the offered QoS

S.No.	Composite Web Service	Cost	QoS
1.	TravelC+HotelF+PickupA	16700	0.9825
2.	TravelB+HotelF+PickupA	17100	0.9591
3.	TravelC+HotelF+PickupD	17000	0.8306
4.	TravelB+HotelF+PickupD	17400	0.8242
5.	TravelG+HotelF+PickupA	17100	0.8167
6.	TravelC+HotelF+PickupE	17000	0.7953
7.	TravelG+HotelF+PickupE	17400	0.7953
8.	TravelG+HotelF+PickupB	17100	0.7909
9.	TravelC+HotelF+PickupB	16700	0.7909
10.	TravelB+HotelF+PickupE	17400	0.7897

In Table 4 a comparison of the number of sets generated for composition with and without proposed approach is presented. For different requirements related to various services, the number of services discovered from the registry is very large as shown in column 4 of the table. Based on the QoS values calculated by the system for different services and cost requirement given by the user, the proposed approach presents with only top ten services to the user that shall fulfill his requirement most suitably.

Table 4. Comparison of Number of Service Sets Generated For With and Without Proposed Method

Travel Service	Hotel Service	Pickup Service	Total Number of Generated Service Sets		Improvement %
			Without Proposed Method	With Proposed Method	
Source :Bhopal Destination: Gwalior Travel Mode: Car AC Required: AC No. of Seats: 4 DateofJ: 15-06-2011	City :Gwalior HotelType:2-star AC Required : AC Single/Dou:Double No. of Rooms : 2 DateF:15-06-2011 DateT: 17-06-2011	City: Gwalior VehicleNa :Qualis AC Req : AC No of Rooms : 1 DateF :15-06-2011 DateT :17-06-2011	72	10	86.11%
Ready to pay : Atmost 17700 units					
Source :Indore Destination: Gwalior Travel Mode: Bus AC Required: NAC No. of Seats: 2 DateofJ: 22-06-2011	City : Gwalior HotelType:3-star AC Required:NAC Single/Dou :Single No. of Rooms : 1 DateF:22-06-2011 DateT: 22-06-2011	City: Gwalior VehicleNa: Santro AC Req : AC No of Rooms : 1 DateF :22-06-2011 DateT:22-06-2011	90	10	88.88%
Ready to pay : Atmost 2700 units					
Source : Gwalior Destination: Indore Travel Mode: Car AC Required: AC No. of Seats: 4 DateofJ: 23-06-2011	City : Indore HotelType:Normal AC Required : AC Single/Dou:Double No. of Rooms : 2 DateF:23-06-2011 DateT: 24-06-2011	City : Indore VehicleNa :Santro AC Req : AC No of Rooms : 1 DateF :23-06-2011 DateT:24-06-2011	144	10	93.05%
Ready to pay : Atmost 13100 units					
Source : Gwalior Destination: Bhopal Travel Mode: Bus AC Required: NAC No. of Seats: 2 DateofJ: 27-06-2011	City : Bhopal HotelType :2-star ACRequired :NAC Single/Dou:Double No. of Rooms : 1 DateF:27-06-2011 DateT: 29-06-2011	City : Bhopal VehicleN:TataIndica AC Req :NAC No of Rooms : 1 DateF :27-06-2011 DateT:29-06-2011	108	10	90.74%
Ready to pay : Atmost 6600 units					

Table shows that the method proposed in this paper is highly useful to the users as only the most suitable service sets are shown to the user.

5 Conclusion

The approach presented in this paper is useful to obtain a complex service constituted by combining the component services that best match to each other with regard to the QoS. The proposed method enables the user to avail a complex service that best meets QoS and cost related requirements. The method is highly useful as it presents very few service sets to the user that have high values for the important QoS parameters namely availability, throughput and response time as evident from the results.

The future work is to devise a mechanism for dynamic selection and composition of services by checking compatibility among component services based on other QoS parameters.

References

1. W3C, Web Services Architecture (2006),
<http://www.w3.org/TR/2004/NOTE-ws-arch-20040211/>
2. Mahmoud, Q.H.: Service Oriented Architecture (SOA) and Web Services: The Road to Enterprise Application Integration (EAI), Oracle (April 2005),
<http://www.oracle.com/technetwork/articles/javase/soa-142870.html> (accessed on October 21, 2011)
3. Wang, Q.-M., Tang, Y., Zhang, Z.-B.: Research In Enterprise Applications of Dynamic Web Service Composition Methods and Models. In: Preceeding of Second International Symposium on Electronic Commerce and Security, pp. 146–150. IEEE (2009)
4. Senkul, P.: Composite Web Service Construction by Using a Logical Formalism. In: Preceeding of 22nd International Conference on Data Engineering Workshops (ICDEW 2006), pp. 56–65. IEEE (2006)
5. Boumhamdi, K., Jarir, Z.: Yet Another Approach for Dynamic Web Service Composition. In: Preceeding of International Conference on Internet Technology and Secured Transactions (ICITST 2009), pp. 1–5. IEEE (2009)
6. Ye, G., Wu, C., Yue, J., Cheng, S., Wu, C.: A QoS-aware Model for Web Services Discovery. In: Preceeding of First International Workshop on Education Technology and Computer Science, pp. 740–744. IEEE (2009)
7. Al-Masri, E., Mahmoud, Q.H.: ‘QoS-Based Discovery and Ranking of Web Services. In: Preceeding of 16th International Conference on Computer Communications and Networks (ICCCN 2007), pp. 529–534. IEEE (2007)
8. QoS for Web Services: Requirement and Possible Approaches, w3c Working Group0 (November 25, 2003),
<http://www.w3c.or.kr/kr-office/TR/2003/ws-qos/#qos-intro>
9. Thirumaran, M., Dhavachelvan, P., Abarna, S., Aranganayagi, G.: Architecture for Evaluating Web Service QoS Parameters Using Agents. Preceeding of International Journal of Computer Applications 10(4), 15–21 (2010)
10. Mahmoud, Q.H.: The QWS Dataset.,
<http://www.uoguelph.ca/~qmahmoud/qws/index.html>

Classification of Rock Textures

Thiagarajan Harinie, I. Janani Chellam, S.B. Sathya Bama,
S. Raju, and V. Abhaikumar

Thiagarajar College of Engineering,
Madurai
harinie.thiagarajan@gmail.com,
jananelangovan@ymail.com,
sbece@tce.edu

Abstract. This paper presents a novel method for the classification of rocks into the three major categories, namely, igneous, sedimentary and metamorphic. Each of these rock types has various sub-types too. The various Tamura Features are formulated and calculated from the input image. The values obtained are compared with the query image by Sum of Squared Distance (SSD). The classified results are then compared with those results of Grey Level Cooccurrence Matrix (GLCM), Color cooccurrence matrix and Moments. The proposed method outperforms the other previously developed methods by providing the classification accuracy of more than 87% for all the three types of rocks. The proposed method significantly improves efficiency with less computational complexity.

Keywords: Color cooccurrence matrix, computational complexity, Grey Level Cooccurrence Matrix (GLCM), Moments, Sum of Squared Distance, Tamura Features.

1 Introduction

Texture is an important feature for any type of image in application and it can be used for image segmentation. Different types of textures were identified. Texture analysis is important in many applications of computer image analysis for classification or segmentation of images based on local spatial variations of intensity or color. A successful classification or segmentation requires an efficient description of image texture.

Rocks are generally classified by mineral and chemical composition, by the texture of the constituent particles and by the processes that formed them. These indicators separate rocks into igneous, sedimentary, and metamorphic. They are further classified according to particle size. The transformation of one rock type to another is described by the geological model called the rock cycle.

Igneous rocks are formed when molten magma cools and are divided into two main categories: plutonic rock and volcanic. Plutonic or intrusive rocks result when magma cools and crystallizes slowly within the Earth's crust (example granite), while

volcanic or extrusive rocks result from magma reaching the surface either as lava or fragmental ejecta (examples pumice and basalt) .

Sedimentary rocks are formed by deposition of either clastic sediments, organic matter, or chemical precipitates (evaporites), followed by compaction of the particulate matter and cementation during diagenesis. Sedimentary rocks form at or near the Earth's surface. Mud rocks comprise 65% (mudstone, shale and siltstone); sandstones 20 to 25% and carbonate rocks 10 to 15% (limestone and dolostone).

Metamorphic rocks are formed by subjecting any rock type (including previously formed metamorphic rock) to different temperature and pressure conditions than those in which the original rock was formed. These temperatures and pressures are always higher than those at the Earth's surface and must be sufficiently high so as to change the original minerals into other mineral types or else into other forms of the same minerals (e.g. by re-crystallization).

The three classes of rocks — the igneous, the sedimentary and the metamorphic — are subdivided into many groups. There are, however, no hard and fast boundaries between allied rocks. By increase or decrease in the proportions of their constituent minerals they pass by every gradation into one another, the distinctive structures also of one kind of rock may often be traced gradually merging into those of another. Hence the definitions adopted in establishing rock nomenclature merely correspond to selected points (more or less arbitrary) in a continuously graduated series.

Rock is a natural texture. Analysis of rock texture has become quite demanding due its varied industrial applications. Rock textures are non homogeneous unlike Brodatz textures [1]. Each rock type has its own nature and its application. Also the granular size, texture and color varies with each type. Already the rock types with iron ore deposit were classified using the digital image analysis technique. The image acquisition and analysis of blasted rocks were conducted in a laboratory for six different rock types. Due to these factors classification of rock texture becomes difficult. Textures can generally be defined either by texture intensity or spectral properties.

Haralick et al was the first to classify images based on textures [2]. Fourteen features were extracted which proved to be computationally expensive. Later Gray level co-occurrence matrix (GLCM) developed and was used for rock classification [3]. It was comparatively better than the previous methods. But still it was less efficient since it didn't take into account the color factor. Cooccurrence matrix was then extended to the color features to get a better retrieval [4].

This paper is to focus both on textural features. In our approach, each texture was subjected to the Tamura Feature and the retrieval of textures is based on the features extracted. This approach yields convincing results, effective than the previous methods.

The rest of the paper is organized as follows: An overview of classification method is given in section 2. Section 3 gives the experimental results to illustrate the efficiency of the algorithm. Section 4 presents experimental results compared to 2 other methods. Section 5 concludes the paper.

2 Methodology

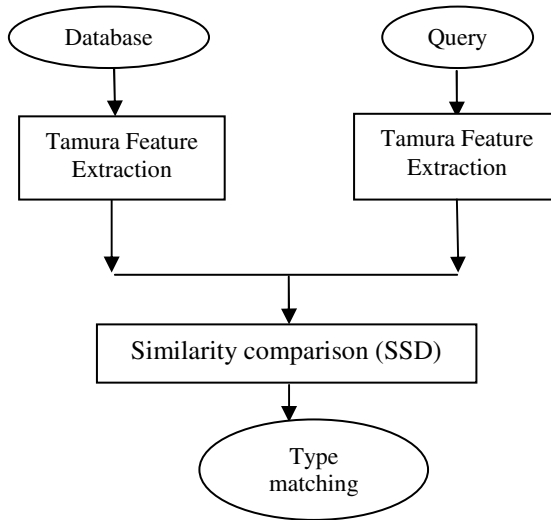


Fig. 1. Flow chart of the proposed method

3 Proposed Method

3.1 Tamura Feature Extraction

The relative brightness of pairs of pixels is computed such that degree of contrast, regularity, coarseness and directionality may be estimated [5]. However, the problem is in identifying patterns of co-pixel variation and associating them with particular classes of textures such as *silky*, or *rough*.

Tamura et al. (1978) proposed a texture representation based on psychological studies of human perceptions. Each texture image in the database is represented as six Tamura features including coarseness, contrast, directionality, line-likeness, regularity, and roughness, to describe low level statistical properties of textures.

Tamura features are visually meaningful, whereas some of CM-features (e.g., entropy and inertia) may not. This advantage makes Tamura features very attractive in texture-based image retrieval

- Coarseness

It refers to texture granularity, that is, the size and number of texture primitives. A coarse texture contains a small number of large primitives, whereas a fine texture contains a large number of small primitives. Coarseness (f_{crs}) can be computed as follows.

$$f_{crs} = \frac{2^k}{n^2} \sum_i^n \sum_j^n p(i, j) \quad (1)$$

- Contrast

It stands for image quality in the narrow sense; it refers the difference in intensity among neighboring pixels. A texture on high contrast has large difference in intensity among neighboring pixels, whereas a texture on low contrast has small difference.

$$f_{con} = \frac{\alpha}{(\mu_4/\sigma^4)^{1/4}} \tag{2}$$

- Directionality

Directionality is a global property over a specific region; it refers the shape of texture primitives and their placement rule. A directional texture has one or more recognizable orientation of primitives, whereas an isotropic texture has no recognizable orientation of primitives.

$$f_{dir} = 1 - r.n_p \sum_p^n \sum_{\phi \in w_p} (\phi - \phi_p). H_D(\phi) \tag{3}$$

- Line-likeness

Line-likeness refers only the shape of texture primitives. A line-like texture has straight or wave-like primitives whose orientation may not be fixed. Often the line-like texture is simultaneously directional.

$$f_{lin} = \frac{\sum_i^n \sum_j^n P_{Dd(i,j)} \cos\left[\frac{(i-j)2\pi}{n}\right]}{\sum_i^n \sum_j^n P_{Dd(i,j)}} \tag{4}$$

- Regularity

Regularity refers to variations of the texture-primitive placement. A regular texture is composed of identical or similar primitives, which are regularly or almost regularly arranged. An irregular texture is composed of various primitives, which are irregularly or randomly arranged (Haralick, 1982).

$$f_{reg} = 1 - r(\sigma_{crs} + \sigma_{con} + \sigma_{dir} + \sigma_{lin}) \tag{5}$$

- Roughness

It refers tactile variations of physical surface. A rough texture contains angular primitives, whereas a smooth texture contains rounded blur primitives.

$$f_{rgh} = f_{crs} + f_{con} \tag{6}$$

The various Tamura Features have their Linguistic terms varying between the two extremes. This variation can be determined from the value of the Tamura Feature calculation. Depending on the variation, the rock classification can be done accurately.

The goal of texture description is to interpret an unknown textures as linguistic terms, that is, as degrees of appearance for each Tamura features. Moreover, membership values of Tamura feature in a linguistic term determines the availability of the form for the feature. For each Tamura feature, membership values of the five linguistic terms are computer by using the term set in the fuzzy clustering. This value depicts the availability of the feature in the Tamura set term. This method is very useful in texture based image retrieval. The effectiveness of the proposed method will be demonstrated through the results.

Table 1. Tamura and its corresponding Linguistic term

Tamura Features	Linguistic Terms				
	Very fine	Fine	Medium coarse	Coarse	Very coarse
Coarseness	Very fine	Fine	Medium coarse	Coarse	Very coarse
Contrast	Very low	Low	Medium contrast	High	Very high
Directionality	Very isotropic	Isotropic	Medium directional	Directional	Very directional
Line-likeness	Very blob-like	Blob-like	Medium line-like	Line-like	Very line-like
Regularity	Very irregular	Irregular	Medium regular	Regular	Very regular
Roughness	Very smooth	Smooth	Medium rough	Rough	Very rough
Coarseness	Very fine	Fine	Medium coarse	Coarse	Very coarse

3.2 Experimental Procedure

The rock images contained in the database are of size 256 X 256. The training set was formed from the database with the classified rocks in the collection of rocks. The three types of rocks namely igneous, sedimentary and metamorphic are considered. An image for each type is standardized. Then the query image is compared with these standardized images and sorted.

Table 2. Count for database and training set

Database characteristics	Database	Training
Number of images	60	50
Image size	256 X 256	256 X 256

4 Results and Discussion

For the experiment, a set of about 100 images were analyzed. Figure 2 shows the sample image set used for the experiment. The sample dataset contains rock images that belong to different types. Each of these images is compared with those of the query set and Sum of Squared Distance (SSD) is calculated.

The obtained results are compared with those of Color cooccurrence Matrix, Grey Level Cooccurrence Matrix (GLCM) and Moments. It can be seen that better classification accuracy can be obtained from that of Tamura Feature Classification.

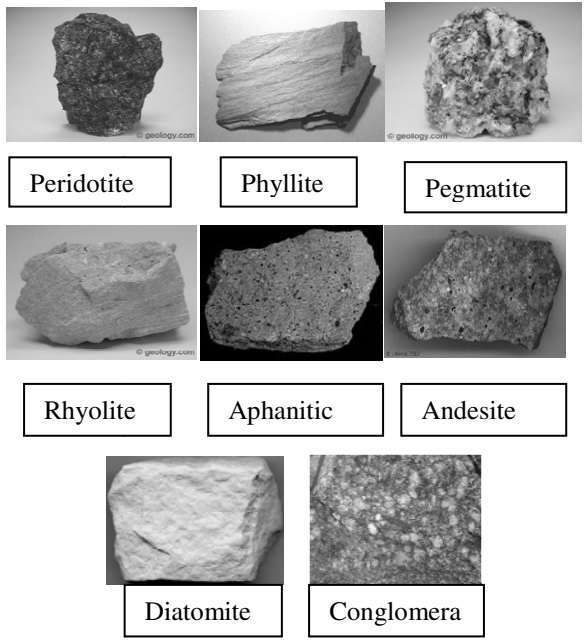


Fig. 2. Sample database of the rock images rock types using the color cooccurrence matrix

Table 3. Sample images and its corresponding Tamura Features


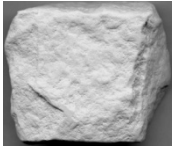
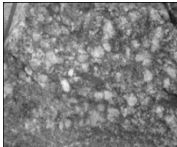
Sl. No	Image	Tamura feature values		Fuzzy Results
1		$fcrs = 0.5284$ $fcon = 0.0013$ $flin = 0.5126$	$frgh = 0.5297$ $freg = 1$ $fdir = 0.5512$	<i>Coarse</i> <i>Rough</i> <i>Isotropic</i>
2		$fcrs = 0.3283$ $fcon = 0.0026$ $flin = 0.4120$	$frgh = 0.3310$ $freg = 1$ $fdir = 1.0443$	<i>Directional</i> <i>Rough</i> <i>Coarse</i>
3		$fcrs = 0.2163$ $fcon = 0.0032$ $flin = 0.3374$	$frgh = 0.4546$ $freg = 1$ $fdir = 1.0314$	<i>Blob like</i> <i>Very rough</i> <i>irregular</i>

Table 4. Sum of Square Distance between the sample rock types using Tamura Feature Extraction

Rock sample	Igneous-intrinsic	Igneous-extrinsic	metamorphic	Sedimentary
Periodite	3767.6	137337.3	81378	66178.37
Pegmatite	2386.16	68313.78	7181.68	3117.06
Andesite	3673.33	10.1313	31378.38	6633.17
Aphanite	31686.33	1378.33	17770.7	3137.07
Phyllite	66363.3	30771.36	3636.1	37117.33
Rhyolite	31397.87	60616.89	6338.18	7137.01
Diatomite	170.17	717.13	611.07	133.710
Conglomerate	11137.01	3331.796	1733.78	13.0317

From Table 4, Periodite and Pegmatite have been classified correctly under Igneous Intrinsic while Andesite and Aphanite have been correctly classified under Igneous Extrinsic. Phyllite and Rhyolite belong to Metamorphic Rocks but Tamura has misclassified Rhyolite under Sedimentary. Diatomite and Conglomerate have been correctly classified under Sedimentary.

Table 5. Sum of Square Distance between the sample rock types using color coocurance

Rock sample	Igneous-intrinsic	Igneous-extrinsic	metamorphic	Sedimentary
Periodite	4797.6	132437.4	85328	99128.32
Pegmatite	4489.56	98453.78	2585.68	3152.09
Andesite	3974.33	10.5454	45478.48	6633.12
Aphanite	45989.33	2378.44	52720.7	3132.07
Phyllite	66363.4	30725.49	3936.5	42512.33
Rhyolite	41392.87	60616.89	6338.58	2132.01
Diatomite	520.12	212.54	651.02	183.750
Conglomerate	55532.01	3335.296	1234.28	14.0312

From Table 5, Periodite has been classified correctly under Igneous Intrinsic while Pegmatite, though it belongs to Igneous Intrinsic, has been misclassified under Metamorphic. Andesite and Aphanite have been correctly classified under Igneous Extrinsic. Phyllite and Rhyolite belong to Metamorphic Rocks but Color Coocurance has misclassified Rhyolite under Sedimentary. Diatomite and Conglomerate have been correctly classified under Sedimentary.

Table 6. Sum of Square Distance between the sample rock types using the GLCM

Rock sample	Igneous intrinsic	Igneous-extrinsic	Metamorphic	Sedimentary
Periodite	210.11	885.25	12555.2	2236.218
Pegmatite	1100.58	85328	6322.3	7500.219
Andesite	5822.18	865.58	556.11	3350.12
Aphanite	11100.58	792.478	52336.2	2122.89
Phyllite	8466.88	47785.44	4100.44	39652.218
Rhyolite	11022.11	234515	5722.12	2425.12
Diatomite	888.06	256.12	414.23	88.06
Conglomerate	9924.023	3352.256	23.10	160.024

From Table 6, Periodite and Pegmatite have been classified correctly under Igneous Intrinsic while Andesite, which belongs to Igneous Extrinsic, has been misclassified under Metamorphic. Aphanite have been correctly classified under Igneous Extrinsic. Phyllite belongs to Metamorphic Rocks, but Rhyolite, which belongs to the same class, has been misclassified under Sedimentary. Diatomite has been correctly classified under Sedimentary. But Conglomerate, which belongs to Sedimentary, has been misclassified under Metamorphic rocks.

Table 7. Sum of Square Distance between the sample rock types using moments

Rock sample	Igneous Intrinsic	Igneous-extrinsic	Metamorphic	Sedimentary
Periodite	0	9.00E-08	1.60E-07	4.00E-07
Pegmatite	2.50E-07	4.00E-08	1.00E-08	3.00E-07
Andesite	4.00E-08	0	1.00E-08	5.00E-04
Aphanite	1.00E-08	0	4.00E-08	5.00E-08
Phyllite	2.50E-07	1.00E-08	3.00E-07	0
Rhyolite	2.50E-07	4.00E-08	0	2.100E-08
Diatomite	7.0711E-09	6.102E-09	5.001E-09	2.001E-09
Conglomerate	7.142E-7	5.123E-09	1.00E-09	3.112E-09

From Table 7, Periodite has been classified correctly under Igneous Intrinsic while Pegmatite, which belongs to Igneous Intrinsic, has been misclassified under Metamorphic. Andesite and Aphanite have been correctly classified under Igneous Extrinsic. Phyllite, though belongs to Metamorphic Rocks, has been misclassified under Sedimentary. Rhyolite has been correctly classified under Metamorphic. Diatomite has been correctly classified under Sedimentary. But Conglomerate, which belongs to Sedimentary, has been classified under Metamorphic rocks.

The comparative graph shown in fig.3 gives a clear proof that the classification of rock textures using Tamura Texture Features gives better classification accuracy of more than 87% in comparison with that of the Color Cooccurrence Matrix method, GLCM and moments.

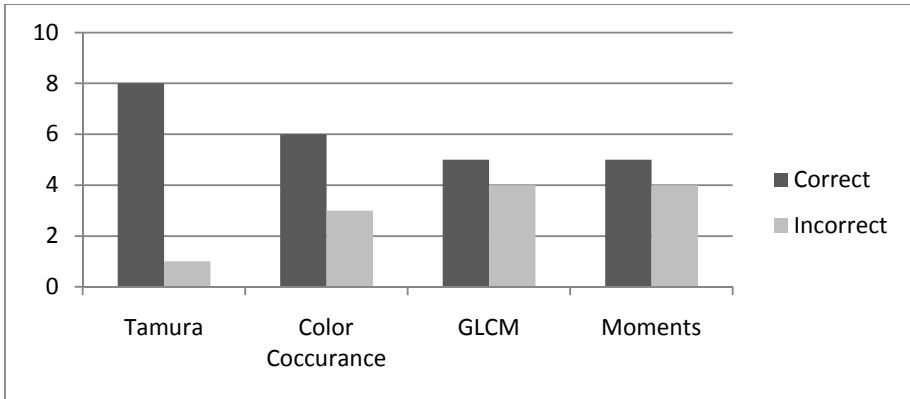


Fig. 3. Comparative graph showing the classification accuracy

4 Conclusion

In this paper an approach to classification of non-homogenous rock textures was proposed using the textural information to classify each type of rock. Compared to the commonly used texture analysis methods which considered only the textures; the Tamura Feature analysis gives better results with more than 87% accuracy. Therefore our approach proved to be useful in classification of non-homogenous rock textures, because the most of the textures occurring in the nature are non-homogenous. These results have practical significance in rock and stone industry. In application area, where rocks and stones have practical significance, the classification methods presented in this paper can be used to classify rock samples into visually similar classes. Because of encouraging results presented in this paper, the idea of testing rock texture samples in different color spaces can be also a subject for further studies.

References

1. Brodatz, P.: Texture, A photographic Album for Artists and Designers. Reinhold, New York (1968)
2. Haralick, R.M., Shanmugam, L., Dinstein: Textural features for image classification. IEEE Trans. Systems. Manufact. Cybernet. 3(6), 610–621 (1973)
3. Partio, M., Cramariuc, B., Gabbouj, M., Visa, A.: Rock texture retrieval using gray level co-occurrence matrix. In: Norsig (October 2002)
4. Partio, M., Cramariuc, B., Gabbouj, M.: Texture retrieval using ordinal co-occurrence features. In: Proceedings of the 6th Nordic Signal Processing Symposium 6th Nordic Signal Processing Symposium - NORSIG 2004, Espoo, Finland, June 9-11 (2004)
5. Lin, H.-C., Chiu, C.-Y., Yang, S.-N.: Finding textures by textual descriptions, visual examples, and relevance feedbacks. Pattern Recognition Letters 24, 2255–2267 (2003)
6. Chatterjee, S., Bhattacharjee, A., Samanta, B., Pal, S.K.: Rock-type classification of an iron ore deposit using digital image analysis technique. International Journal of Mining and Mineral Engineering (2008)
7. Lepisto, L., Kunttu, I., Autio, J., Visa, A.: Rock image classification using non-homogeneous textures and spectral imaging. In: WSCD (February 2003)

Design and Implementation of an Effective Web Server Log Preprocessing System

Saritha Vemulapalli¹ and M. Shashi²

¹ Department of Information Technology,
VNR Vignana jyothi Inst. of Engg. & Tech,
Hyderabad, A.P, India
saritha_vemulapalli@yahoo.com

² Department of CS & SE,
Andhra University College of Engg (A),
Visakhapatnam, A.P, India
smogalla2000@yahoo.com

Abstract. WWW constitutes huge repository, distributed and dynamically growing hyper medium, supporting access to information and services. As more organizations rely on WWW to conduct business, user behavior analysis becoming difficult in web-based applications. Information about user's interactions with website is stored in server logs and serves as huge electronic survey of website. Web usage mining deals with discovering usage patterns from server logs in order to understand and better serve the needs of web users. The raw information contained in log file represents noisy data. Preprocessing includes cleaning, user identification, sessionization, path completion & structurization and is a prerequisite for improving accuracy and efficiency of the subsequent mining process. This paper emphasizes on an effective web log preprocessing system. Experimental results proved that the proposed system reduces the size of log file down to 12% and improves the performance of preprocessing in identifying users, sessions, path completion and structurization.

Keywords: Data Mining, Web Log Mining, Web Usage Mining, Preprocessing, Cleaning, User Identification, Sessionization, Path Completion.

1 Introduction

Since 1991, WWW became so popular & has a rapid development. Now it has formed a great distributed information source including 8.75 millions websites, 2.5 billions web pages and great many users [1]. The WWW constitutes a huge repository, widely distributed and dynamically growing hyper medium, supporting access to information and services. With the explosive growth of information sources available on the WWW, providing web users with more exactly needed information is becoming a critical issue in web-based applications. It has become necessary for users to utilize automated tools in order to find, extract, filter, and evaluate the desired information. As a result, web usage mining has attracted lot of attention in recent time [2].

Web-based applications generate and collect large volumes of data in their day-to-day activities. Majority of this data is generated automatically by web servers and collected in server logs in an unstructured format. Web mining is the application of data mining which deals with the extraction of interesting knowledge from the WWW documents and services which are expressed in the forms of textual, linkage or usage information [3]. Web mining can be divided into web content mining, web structure mining and web usage mining. Web content mining is the process of discovering useful knowledge from the raw data (text, image, audio or video data) available in web pages. Web structure mining is the process of analyzing the link between pages of a web site using web topology. Cooley et al. [4] introduced the term web usage mining in 1997 and is defined as process of extracting useful information from server logs (i.e. user's history) to improve web services and performance. Obtained user access patterns can be used in variety of applications, such as to identify the typical behavior of the users [5], making clusters of users with similar access patterns and by adding navigational links [6]. Typical applications are website design & management, web personalization, adaptive websites, recommendation systems, cross marketing strategies, promotional campaigns and user behavior analysis.

The paper is organized as follows. Section 2 describes overview of web usage mining. Design & implementation of proposed preprocessing system and also related algorithms are presented in section 3. Section 4 covers experimental results, proves the effectiveness & efficiency of our algorithms. Conclusions are in section 5.

2 Web Usage Mining Process

Web usage mining is the discovery of user access patterns from server logs, consists of data collection, preprocessing, pattern discovery & analysis and visualization [7]. The data which is used for mining process can be collected from server side, client side, proxy server, website topology, web page contents & user profile information. Server logs are the primary source of data for web usage mining that are collected as a result of users interactions with website, represented in standard formats (e.g. Common Log Format [8] and Extended Common Log Format [9]). The raw information in a web server log file doesn't represent a structured, complete, reliable & consistent data. Preprocessing techniques can improve the quality of the data involves cleaning, user identification, session identification, path completion and data structurization [10]. Statistical & data mining techniques can be applied to the preprocessed web log data, in order to discover statistics & user access patterns and are represented using visualization techniques such as charts, graphs & reports.

2.1 Common Log Format

Each line in a log file represented in the common log format has the following syntax. [Host/IP Rfcname Userid [DD/MMM/YYYY: HH:MM:SS -0000] "Method /Path HTTP/1.x" Code Bytes]

A "-" in a field indicates missing data.

2.2 Extended Common Log Format

It's an extension to common log format, having some additional information like user_agent, cookie and referrer. User_agent is the visitor's browser version & O.S. Referrer defines the URL from where the visitor came from. Each line in a log file represented in the extended common log format has the following syntax.

```
[s-computername s-ip s-port c-ip rfcname cs-userid date time cs-method cs-uri-stem cs-uri-query cs-version sc-status time-taken sc-bytes cs(user-agent) cs(cookie) cs(referrer)]
```

3 Design and Implementation of Proposed Preprocessing System

The proposed preprocessing system uses server logs of www.vnrvjiet.ac.in, is an implicitly generated data as a result of user interactions with a website are represented in Extended common log format (ECLF). Most of the researchers considered web server log file as most reliable and accurate for WUM process.

The following are some of the drawbacks of using server logs.

- Since HTTP is stateless, web server logs do not identify sessions or users.
- Web cache keeps track of pages that are requested and saves a copy of these pages for a certain period. Hence, these requests are not recorded in log files.
- IP address misinterpretation due to shared computers.
- The browser's back button is "the second most used feature" on the web; it accounts for 41% of all user interaction requests for web documents [11].

Our proposed system addresses all the above issues.

3.1 Data Preprocessing

As the web server logs are not designed for data mining, preprocessing must be carried out in order to obtain reliable and accurate data. Low-quality data will lead to low-quality mining results. Data preprocessing techniques can improve the quality of the data, thereby helping to improve the accuracy and efficiency of the subsequent mining process. Nearly 80% of mining efforts are required to improve the quality of data [12]. The proposed preprocessing system consists of components such as cleaning, user identification, session identification, path completion and data structuring is shown in Fig. 1. The implementation issues are explained below.

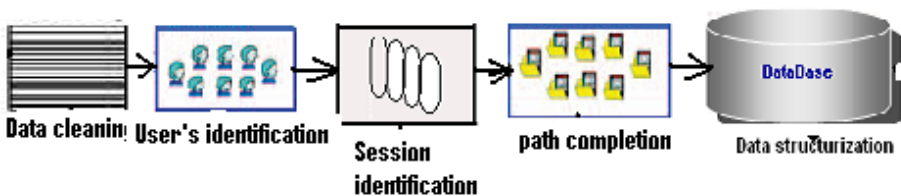


Fig. 1. Proposed Preprocessing System

Data Cleaning: The process of removing entries which are irrelevant and redundant in pattern discovery. HTTP is a stateless & connectionless protocol which requires separate connections for every file requested from the web server. In general a user does not explicitly request all of the graphics on a web page, which are automatically downloaded due to the embedded HTML tags. In the real world data, irrelevant files are found up to a ratio of 10:1, depending on how many graphics and other files the web pages contain [10]. The main intent of web usage mining is to get a picture of the user's behavior, other than file requests that the user did not explicitly request. Removing such entries decreases the memory usage and improves the performance.

The following rules are used for data cleaning in our proposed system:

- i) Removing all the attributes which contain no data at all and are not essential for the analysis.
- ii) Removing log entries covering image, sound, video, flash animations, frames, pop-up pages, script's and style sheet files.
- iii) Removing access records generated by automatic search engine agents such as crawler, spider, robot, etc. Spiders are widely used in web search engine tools to update their search indexes [13]. Spider requests can be identified by looking
 - a) All hosts that have requested the page "robots.txt."
 - b) Many crawlers voluntarily declare themselves in user agent field of log, by referring user agent field whether it contains either a URL or an email address.
- iv) Removing log entries that have status of "error" or "failure". All the entries with a status code other than the 200 range are removed.
- v) Removing log entries that have http request method other than Get or Post.

The following is an algorithm for data cleaning:

Input: Records of server log file, which is represented as log_file $R = \{R_1, R_2, \dots, R_i, \dots, R_n\}$. Where $R_i = \langle F_1, F_2, \dots, F_j, \dots, F_n \rangle$ is a record in log_file and is defined as $\langle s\text{-computername}, s\text{-ip}, s\text{-port}, c\text{-ip}, rfcname, cs\text{-userid}, date, time, cs\text{-method}, cs\text{-uri-stem}, cs\text{-uri-query}, cs\text{-version}, sc\text{-status}, time\text{-taken}, sc\text{-bytes}, cs(user\text{-agent}), cs(cookie), cs(referrer) \rangle$.

Output: log_information & data_cleaning, are database object's.

Algorithm:

Begin

1. Remove non essential attributes for the analysis such as $\langle s\text{-computername}, s\text{-ip}, s\text{-port}, rfcname, cs\text{-uri-query}, time\text{-taken}, sc\text{-bytes} \rangle$ from log_file.
 2. Remove the attributes which doesn't contain data in all records of log_file. // indicates missing values.
 3. FOR each Record R_i in log_file // $1 \leq i \leq n$
 - DO Insert R_i into log_information
- END FOR

```

4. FOR each Record  $R_i$  in log_information
    DO IF( $R_i$  doesn't represent image, sound, video, flash
        animation, frame, pop-up page, script, style
        sheet file, crawler request, error request and
        other than get or post request) Then
        Insert  $R_i$  into data_cleaning
    END IF
END FOR
END

```

User Identification: The process of identifying the unique users, who is interacting with a website using the web browser. The analysis of web usage doesn't require knowledge about a user's identity. However it is necessary to distinguish among different users.

The following rules are used for user identification in our proposed system:

- i) If the IP address is different is assumed as new user.
- ii) If the IP address is same, but with different operating system or browser software is assumed as new user.
- iii) If the IP address, operating system and browser software are same, but with different http version is assumed as new user.

The following is an algorithm for user identification:

Input: Records of data_cleaning "R" = $\{R_1, R_2, \dots, R_i, \dots, R_n\}$.

Output: Records of users_info "U" = $\{U_1, U_2, \dots, U_j, \dots, U_n\}$, is a database object. Where $U_j = \langle F_1, F_2, \dots, F_k, \dots, F_n \rangle$ is a record in users_info.

Algorithm:

```

Begin
1. Select IP, user_agent, version fields of records of
   data_cleaning.
2. Insert  $R_1$  into users_info //  $R_1$  is a first record in R
3. FOR each Record  $R_i$  in data_cleaning //  $1 \leq i \leq n$ 
    DO FOR each Record  $U_j$  in users_info
        IF((IP is different ) OR (IP is same, but with
            Different operating system or browser
            software) OR (IP, operating system and
            browser software are same, but with different
            http version)) Then
            Insert  $R_i$  into users_info ;
        END IF
    END FOR
END FOR
END

```


User’s Session Identification : Web log span long periods of time; it is very likely that users will visit the web site more than once. The process of identifying sequence of activities of a single user during a single visit at a defined duration [14]. Since HTTP protocol is stateless and connectionless discovering the user’s sessions from server log is a complex task.

The following rules are used for session identification in our proposed system:

- i) A new session begins each time when there is a new user.
- ii) A new session begins each time when the time gap between consecutive requests made by the same user exceeds threshold $\Delta t=10$ minutes when the referrer is “-”.
- iii) A new session is identified if the URL in the referrer field has never been accessed before in a current session.

Path Completion: Some important page requests are not recorded in server log due to the cache, thus causing the problem of incomplete path. It is the process of reconstructing the user’s navigation path, by appending missed page requests (page requests that are not recorded in server log) within the identified sessions.

The following rules are used for path completion in our proposed system:

- i) With in the identified user’s sessions, if the URL in the referrer field of the page request made is not equivalent to URL of last page user has requested & if the URL in the referrer field is in the user’s history, it is assumed that user uses “back” button. Missing page references that are inferred through this rule are added to the user’s session file.

The following is an algorithm for sessionization & path completion:

Input: Records of data_cleaning “R” = $\{R_1, R_2, \dots, R_i, \dots, R_n\}$ sorted in ascending order of date, time and Records of users_info “U” = $\{U_1, U_2, \dots, U_j, \dots, U_n\}$.

Output: Records of users_sessions “S” = $\{S_1, S_2, \dots, S_k, \dots, S_n\}$, is a database object. Where $S_k = \{U_j, path_i\}$ is a session in S, U_j is a record in users_info & $path_i$ is defined as $url_{i1} \rightarrow url_{i2} \rightarrow \dots \rightarrow url_{in} // 1 \leq i \leq n$ And

Records of users_sessions_path “RS” = $\{RS_1, RS_2, \dots, RS_k, \dots, RS_n\}$, is a database object. Where $RS_k = \{U_j, path_i\}$ is a reconstructed session in RS, U_j is a record in users_info & $path_i$ is defined as $url_{i1} \rightarrow url_{i2} \rightarrow \dots \rightarrow url_{in} // 1 \leq i \leq n$

Algorithm:

Begin

Set $S = \{ \}, RS = \{ \};$

FOR each Record U_j in users_info

 Create a new Session S_k & Reconstructed Session RS_k ;

 DO FOR each Record R_i in data_cleaning

 DO IF (Values of IP, user_agent & version are same) Then

 DO IF ((Referrer is ‘-’ & Time gap between consecutive requests by the same user > 10min)

 OR (URL in Referrer field has never been

```

        accessed before in current session)) Then
        Create new Session  $S_k$  & Reconstructed Session  $RS_k$ ;
        Add uri-stem field to  $path_i$  of the current
        Session  $S_k$  &  $path_i$  of the current Reconstructed
        Session  $RS_k$ ;
    Else
        Add uri-stem field to  $path_i$  of the current
        Session  $S_k$  ;
        IF(URL in Referrer field is not equivalent to
        URL of last page user has requested) Then
            Add missing page references to  $path_i$  of the
            current Reconstructed Session  $RS_k$  ;
        Else
            Add uri-stem field to  $path_i$  of the current
            Reconstructed Session  $RS_k$  ;
        END IF
    END FOR
FOR each Session in  $S_k$  &  $RS_k$ 
    DO Insert  $S_k$  into users_sessions;
    Insert  $RS_k$  into users_sessions_path;
END FOR
END FOR
END

```

Data Structurization: The process of transforming and storing the data into suitable form for input to the pattern discovery. Different tables are designed in the relational database for each object, identified in various stages of preprocessing.

4 Experimental Results

The proposed system was developed based on IIS web server log represented in ECLF, using java programming language. Experimental analysis is carried out to validate the effectiveness and efficiency of the proposed preprocessing system. The server log file of www.vnrvjiet.ac.in of 15th Nov 2010, having 10,375 records is selected for analysis. The results of preprocessing are shown in Table 1. After cleaning the No. of records reduces down to 1,220 (12% of original records), 235 unique users & 589 user's sessions are identified.

Table 1. The Results of Data Preprocessing

Records in logfile	Records after cleaning	NO of unique Users	Sessions
10,375	1,220	235	589

Table 2 shows the results of data cleaning process. 1 represents records in raw server log file, 2 represents records after removing image, sound, video, flash animations, frames, pop-up pages, script's and style sheet files. 3 represents records after further removing crawler requests. 4 represents records after further removing error requests. Table 3 shows the results of user identification. 1 represents unique users identified using IP address, 2 represents unique users identified using IP address & user_agent. 3 represents unique users identified using IP address, user_agent & version.

Table 2. Results of Data Cleaning Process

Cleaning Process	No. of Records
1	10,375
2	1581
3	1230
4	1220

Table 3. Results of User Identification Process

User identification process	No. Of users
1	215
2	234
3	235

Table 4. Results of Session Identification Process

Session Identification Process	No. of sessions
1	269
2	253
3	818
4	589

User Identification using the IP address alone is not sufficient and reliable. This can result in several users being erroneously grouped together as one user. Because although an IP address may represent one person only, an IP address is in most cases shared by more than one person (at a library, internet cafe or one user uses multiple computers). So, different users sharing the same host can not be distinguished.

Our experimental results proved that unique users can be identified more effectively using user_agent & version field's along with IP address. The rationale behind this rule is that a user, when navigating the web site, rarely employs more than one browser, much more than one OS.

Table 4 shows the results of user's session's identification process. 1 represents sessions identified based on time gap between two consecutive page requests exceeds 10min's. 2 represents identified sessions based on session duration as 30min's. 3 represents identified sessions based on if the URL in the referrer field has accessed before in a current session. 4 represents identified sessions based on proposed session identification rules.

Time based methods are not reliable because users may involve in some other activities after opening the web page and factors such as busy communication line, loading time of components in web page, content size of web pages are not considered. Referrer based method introduces the confusion when user types URL directly or uses bookmark to reach pages not connected via links and identified sessions may contains more than one visit by the same user at different time. Our experimental results proved that session's can be identified more effectively using our proposed session identification rules.

5 Conclusion and Future Enhancements

The raw information contained in a web server log file as a result of user's interactions with a website doesn't represent a structured, complete, reliable & consistent data. As the web server logs are not designed for data mining, preprocessing must be carried out to improve the accuracy and efficiency of the subsequent mining process. Low-quality data will lead to low-quality mining results. Server logs of www.vnrvjiet.ac.in are analyzed using the proposed preprocessing system in order to identify unique users, user sessions & path completion and data structurization, which play a major role in web usage mining process in order to

discover useful hidden patterns reflecting the typical behavior of users. Experimental results proved that the proposed system reduces the size of log file down to 12% and improves the performance of preprocessing in identifying users, sessions, path completion and structurization.

The proposed system can be enhanced in future for more accurate session identification & path completion.

References

1. Gudivada, V.N.: Information retrieval on the World Wide Web. *IEEE Internet Computing* 1(5), 58–68 (1997)
2. Cooley, R., Mobasher, B., Srivastava, J.: Web mining: information and pattern discovery on the World Wide Web. In: *International Conference on Tools with Artificial Intelligence*, pp. 558–567. IEEE, Newport Beach (1997)
3. Cooley, R., Mobasher, B., Srivastava, J.: Data preparation for mining World Wide Web browsing patterns. *Journal of Knowledge and Information System*, 1–27 (1999)
4. Cooley, R., Mobasher, B.S., Srivastava, J.: Grouping Web page references into transactions for mining World Wide Web browsing patterns. In: *Knowledge and Data Engineering Workshop*, pp. 2–9. IEEE, New port Beach (1997)
5. Srivastava, J., Cooley, R., Deshpande, M., Tan, P.-N.: Web usage mining: Discovery and applications of usage patterns from Web data. *SIGKDD Explorations* 1, 12–23 (2000)
6. Masegla, F., Poncelet, P., Teisseire, M.: Using data mining techniques on Web access logs to dynamically improve Hypertext structure. *ACM SigWeb Letters* 8(3), 13–19 (1999)
7. Pabarskaite, Z., Raudys, A.: A process of knowledge discovery from web log data: Systematization and critical review. *Journal of Intelligent Informatin Systems* 28(1), 79–104 (2007)
8. Configuration file of W3C httpd (1995),
<http://www.w3.org/Daemon/User/Config/>
9. W3C Extended Log File Format (1996),
<http://www.w3.org/TR/WD-logfile.html>
10. Cooley, R., Mobasher, B., Srivastava, J.: Data Preparation for Mining World Wide Web Browsing Patterns. *J. Knowledge and Information Systems* 1(1), 5–32 (1999)
11. Catledge, L.D., Pitkow, J.E.: Characterizing browsing strategies in the world-wide web. *Computer Networks and ISDN Systems* 27, 1065–1073 (1995)
12. Frieder, O., Grossman, D.A.: *Information Retrieval: Algorithms and Heuristics*. The Information Retrieval Series, 2nd edn (2004)
13. Tanasa, D., Trousse, B.: Advanced data preprocessing for intersites Web usage mining. *IEEE Intelligent Systems* 19, 59–65 (2004)
14. Spiliopoulou, M.: Managing Interesting Rules in Sequence Mining. In: Żytkow, J.M., Rauch, J. (eds.) *PKDD 1999*. LNCS (LNAI), vol. 1704, pp. 554–560. Springer, Heidelberg (1999)

Component Based Resource Allocation in Cloud Computing

Sumeet S. Vernekar and Pravin Game

Department of Computer Engineering,
PICT, Pune, Maharashtra, India
{sumeet.vernekar, pravingame}@gmail.com

Abstract. In today's world providing on demand computing and storage have become need of time. People are focusing more on the web than that of the local computing due the availability of the portable devices. This has led to the huge demand of Cloud Computing. In today's scenario, many companies are moving towards the cloud for computing and storage resources, as Cloud provides these resources on "pay per use" basis which make it more convenient for the companies to relay on the cloud for the resources. As the demand of the cloud goes on increasing the problem of resource allocation and management will arise. This paper provides a *Component Based Resource Allocation Model* to provide the future resource allocation and management need in cloud computing environment.

Keywords: Cloud Computing, Resource Allocation, Distributed System, Client Server.

1 Introduction

Cloud computing has emerged as a major game changer in today's mobile environment. Cloud is not a new but it is evolved from grid and relies on Grid as infrastructure support. Grid on the other side is project oriented where peoples work on given resource for a particular reason [1], whereas cloud is wide spread. This led to the increased attention towards cloud. The cloud computing system provides people the features like on demand computing and on demand storage. Cloud provides services [1] such as Infrastructure as a Service (IaaS), Software as a Service (SaaS), and Platform as a Service (PaaS). Now from the business centric view, resource availability is of major concern. The companies buy a huge amount of resources, which includes large servers fully equipped with processor consisting of hundreds of cores and thousands of petabytes of storage capacity on lease basis or even some of the companies own a few servers, which leads to higher cost with additional maintenance overhead. An alternative to this is the cloud resources, which provides access to their resource on "pay per use" basis with no overhead of maintenance. As the time will go almost every company will rely on cloud for their resources. In this paper we will focus on resource allocation and management schemes called the *Component based resource allocation model* (CRA), which provides an improved

resource allocation model beneficial for the future cloud computing environment. The paper is organized as follows, in this section 2 related works are studied, in section 3 proposed model is presented, in section 4 the components involved are focused, in section 5 mathematical model is proposed.

2 Related Work

Cloud is considered the most challenging research field in the IT World. The major concern in cloud will is providing good quality of service (QoS) which supports dynamic discovery and reservation of resources for the customer's request. For this an efficient information monitoring system is proposed in GARA architecture [7].

A sufficient amount of work is done on managing the storage resources for the Data-intensive application in the research field in the Data Grid Design [8].

Making the resources available dynamically to the customers over the cloud is the need of the cloud system. An approach called Data Diffusion approach [10] focuses on this aspect. The customers that join the cloud have different requirement for their request which are located at various different locations, so a great challenge is in integrating and coordinating these resources. This issue is covered in the Nimrod/G architecture [11]. Falkon architecture [9] handles the multiple tasks in the cloud.

The concern to provide performance-QoS and economic-QoS in the cloud environment has been considered in the NECDA resource allocation algorithm [3]. The Virtual Machine's (VM) are allocated to the services in the cloud which are created on the physical machine or the node of the cloud system. The Improved Genetic Algorithm proposed in [4] maximizes the resource usage of VM.

In cloud, pricing and allocation of the resource must be done uniformly, so people can request for the resource, without taking in to consideration the price change and available resources. So the solution for this problem has been proposed in [5]. The problem to schedule applications among cloud services that takes both data transmission cost and computation cost into account is overcome by the Revised Discrete Particle Swarm Optimization (RDPSO) [6] model.

A hierarchical P2P scheme is proposed in [2], it provides a high level of abstraction for managing resources in the cloud.

The proposed model, "*Component based resource allocation model*" (CRA) is motivated from the hierarchical P2P scheme [2] and the GARA architecture [7] which focuses on information of the resources which help in managing and allocating resources. It provides a mechanism to provide information regarding each and every resource and also about creating, monitoring and managing the resources independently and uniformly and make this information available centrally.

3 Proposed Model

This model basically is derived by the concept of the *Hierarchical P2P Scheme* [2] which involves the concept of Metascheduler and Superscheduler. As cloud uses grid as its backbone infrastructure, various virtual organizations (VO) will be involved in providing the resource to the cloud. The metascheduler is the node in the VO with

highest configuration. The metascheduler will maintain the information about the node in the VO in a structure called the *Available Node List* (ANL). The superscheduler is one of the metascheduler among all the other metascheduler with the highest configuration among those metascheduler. The decision of who will be the metascheduler among all the nodes in the VO and who will be the superscheduler among all the other metaschedulers is done on the basis of the capacity degree Θ [2]

$$\Theta = \left(\frac{\sum S}{\sum (\Delta S_i)} * \infty \right) + \sum (Pc_n * \varpi)$$

where S_i denotes the amount of services , ΔS_i denotes the average execution time of service i , Pc_n denotes the power computational of resource n , ∞ is related to the amount of service, ϖ is related to the power computational.

The superscheduler maintain all the information about the available resources, the queue for the service’s submitted by the customer, the components responsible for allocation and maintenance and the ANL of the nodes. Entities involved:

1. Superscheduler: The superscheduler is the top level node in the hierarchy and is responsible for the overall allocation and maintenance of the resources. Every request in the cloud first reaches the superscheduler.
2. Virtual Machine (VM): The actual customer’s application is deployed here. Whenever the request is received from the user a VM is created for the request on the physical machine and the resources are allocated. Any VM can be started, stopped and migrated from one physical machine to the other physical machine in the VO depending on the cloud service provider requirement.
3. Physical Machine: These are the computing servers that provide resources for creating the VM.

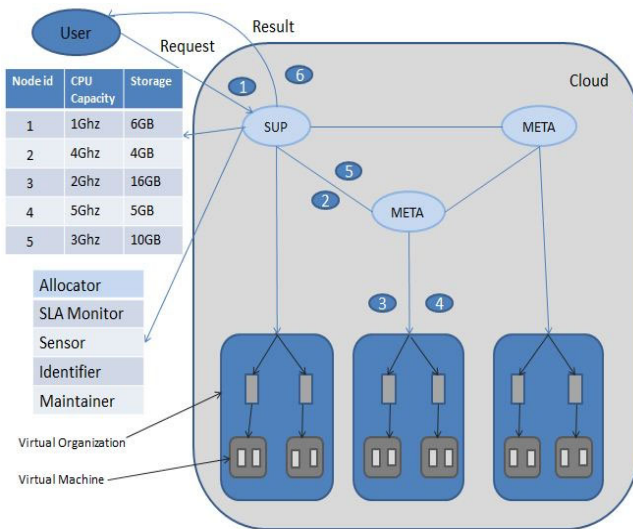


Fig. 1. Component based resource allocation

All the nodes in the VO are connected to each other. These nodes are in turn connected to the other VO via the metascheduler (META).

All the information related to the available resources is stored in a table called as the *Available Resource Table* which is present in the superscheduler (SUP). The information stored in this table is updated periodically. The entries in this table are Node id, Available CPU Capacity and Available Storage as shown in the Fig.1. This table is the array of list with a pointer pointed to the node on which next service request is deployed. This pointer will be adjusted by various components.

4 Components Involved

4.1 Superscheduler

The customer sends the request with the following requirement to the superscheduler.

- I. Throughput (%)
- II. Average Response Time
- III. Application Code
- IV. Operating System

The superscheduler maintains a queue (Q) for storing the requested services. Whenever the request for a service comes to the cloud service provider it generates the components periodically such as the *Allocator*, *SLA Monitor*, *Sensor*, *Identifier* and *Maintainer*. The *Available Resource Table* is empty initially, it is then filled by these components and updated periodically. This information generally contains the current utilization of the resources.

4.2 Allocator

When a request for a service is received by the cloud service provider it will put that request in the queue which is maintained at the superscheduler side. The allocator will then take request from the queue and will allocate the service a virtual machine (VM) with the required CPU capacity, storage and the requested operating system.

The allocation of the VM to the requested service is done on the basis of available resources on the nodes in the VO's. If appropriate resources are available with the requested operating system then the service is deployed over the VM.

If enough resources are not available then service is kept in the queue. The sensor checks for a particular node on the VO which is free or looks for a new node by checking the *Available Node List* of the VO and updates the *Available Resource Table*.

4.3 SLA Monitor

A Service Level Agreement (SLA) [1] is maintained between the cloud service provider and the customer, whenever a service is deployed. The SLA Monitor component maintains and monitors the SLA. SLA is made based upon the average response time and throughput requested by the customer. SLA Monitor keeps track of the SLA by maintaining a variable called as the SLAC. The value of this variable is

calculated based on the resources and the throughput required. Based upon the SLAC value the further action related to the appropriate allocation of resources is been made. Each node will maintain this SLAC variable. The sensor will query each node about its utilization, the nodes will provide the sensor with the SLAC variable value which it updates in the *Available Resource Table*. The value of the SLAC variable are divided into three categories.

1. Normal Load: If the Response time is 10% more than given average response time and 10% more than the throughput, then node is said to have a normal case and SLAC=1.
2. Less Load: If the Response time is 10 to 20 % more than the given average response time or 10 to 20% more than the throughput, then node is said to have less load and SLAC= 2.
3. More Load: If the Response time is 20 % more than the given average response time or 20% more than the throughput, then the node is said to have more load and SLAC=3

The values of the response time and throughput taken above are hypothetical and can be changed according to the cloud service provider's requirement.

4.4 Sensor

The sensor component periodically checks for the utilization information of each node and update the information in the *Available Resource Table*. It also decides whether a node with allocated service has less load or more load based upon the CPU utilization. Based upon this decision, it decides whether to shift the virtual machine (VM) on to another node or not. The decision is based upon the following assumption.

1. If CPU utilization is more than 90%, then select the node with CPU utilization less than 90%. If the next node has SLAC=3 (More load), then see the next node with SLAC=1(Normal load) or 2(Less load) and shift the VM on that node.
2. If the CPU utilization is less than 50% then no need to shift the VM for that node. (Normal load)

The values of CPU utilization taken above are hypothetical and can be changed based upon the cloud service provider's requirement.

4.5 Identifier

The identifier component is responsible for discovery of the newly added nodes in the cloud. Whenever a new node is to be added in the cloud in a VO, a request is send to the metascheduler. The metascheduler will update its *Available Node List* and adds that node to the list and assigns it a unique node id. The identifier will periodically visit each metascheduler and check the *Available Node List*. If it finds a new node added it takes its node id and visit the node, checks its configuration and updates information in the *Available Resource Table*.

4.6 Maintainer

The maintainer checks each node on the cloud by sending each node a request. If each node on the cloud responds the maintainer with a reply, then the node is said to be alive otherwise it is said to be dead node and is removed from the *Available Resource Table* and *Available Node List*. The maintainer also removes VM's with expired SLA.

5 Mathematical Model

Consider a cloud system

$A = \{ \text{Sup, Meta, Comp, T, S, C} \}$

where, $\text{Sup} = \{ \text{ART, ANL, Comp} \}$ – Superscheduler

$\text{Meta} = \{ \text{ANL} \}$ – Metascheduler

$\text{ART} = \{ \text{Nid, P, M} \}$ – Available Resource Table

$\text{ANL} = \{ \text{Nid} \}$ – Available Node List

$\text{Comp} = \{ \text{Id, Sen, All, SLA, Main} \}$ – Components

where, Id – Identifier Component

Sen – Sensor Component

All – Allocator Component

SLA – SLA Monitor Component

Main – Maintainer Component

$T = \{ t_1, t_2, \dots, t_m \}$ – set of resource type

$S = \{ s_1, s_2, \dots, s_n \}$ – set of services to be deployed in the cloud environment

$C = \{ c_1, c_2, \dots, c_n \}$ – set of client workload and respective SLA. $c_i \in C$

where, $c_i = \{ s, \lambda, \rho \}$

where, $s \in S$ is the service been deployed

λ is the workload intensity

ρ is the client requested average response time and throughput

Functions Definition

$V \in [S \rightarrow T]$ specifies which resource types are required by service $s \in S$.

$F \in [S \times T \rightarrow I_{s,t}]$ referred to as resource allocation function assigns to each service $s \in S$ a set of instances $I_{s,t}$ of resource type $t \in T$ (e.g. VM instances).

For $i \in I_{s,t}$, $i = \{ \Pi, P, P', M, M' \}$

where, Π is the processing rate of processing resource

P is the processing resources currently allocated

P' is the maximum number of processing resources that can be allocated

M is the available storage space.

M' is the maximum storage space.

Performance metrics

$X(c)$ is the total number of request of the client workload $c \in C$ completed per unit of time (requested throughput)

$R(c)$ is the average response time of a service request in client workload $c \in C$.

$U(t)$ is the maximum allowed average utilization for resource type $t \in T$ over all instances of the resource

$U'(t)$ is the maximum allowed average utilization for resource type $t \in T$.

Conditions imposed (SLA)

$P_{X(c)}$ for $c \in C$ is defined as $(X(c) \leq C[\lambda])$

$P_{R(c)}$ for $c \in C$ is defined as $(R(c) \leq C[\rho])$

$P_{U(t)}$ for $t \in T$ is defined as $(U(t) \leq U'(t))$

Components involved

1. Allocator Component:
 - if $(\forall c \in C : P_{X(c)} \wedge P_{R(c)}) \wedge (\forall t \in T : P_{U(t)})$ then,
 - $\exists c \in C \wedge s \in S \wedge t \in T$
 - $F(C[s], t) \leftarrow F(C[s], t) \cup i$ where $i \in I_{s,t}$
 - end
2. Identifier Component:
 - if $ANL \leftarrow N'$ then
 - $N' \leftarrow \{N'_{id}, P', M'\}$
 - end
 - $ART \leftarrow N'$ where, N' is a new node in the cloud
3. Sensor Component:
 - if $\exists t \in C[s] : U(t) > 90\% (U'(t))$ then
 - $i[p] \leftarrow i[p]+1$ where $U(t) < 90\% (U'(t))$
 - if $\exists i[p] : SLAC=3$ then
 - $i[p] \leftarrow i[p]+1 \wedge SLAC < 3$
 - end
 - if $\exists i[p] : SLAC=2 \vee SLAC=1$ then
 - $i[p] \leftarrow i[p]$
 - end
 - end
 - if $\exists t \in C[s] : U(t) < 50\% (U'(t))$ then
 - $i[p] \leftarrow i[p]$
 - end
4. SLA Monitor Component:
 - Normal Load:
 - if $\forall c \in C : P_{R(c)} = R(c) + 10\% (R(c)) \vee P_{X(c)} = X(c) + 10\% (X(c))$
 - then $SLAC=1$
 - end
 - Less Load:
 - if $\forall c \in C : P_{R(c)} = R(c) + 10 \text{ to } 20\% (R(c)) \vee P_{X(c)} = X(c) + 10 \text{ to } 20\% (X(c))$ then $SLAC=2$
 - end
 - More Load:
 - if $\forall c \in C : P_{R(c)} = R(c) + 20\% (R(c)) \vee P_{X(c)} = X(c) + 20\% (X(c))$
 - then $SLAC=3$
 - end

5. Maintainer:
 if $\forall c \in C \wedge \exists t \in T \wedge \exists i \in F(C[s], t) : i[p] = 0$ then
 $i[p] \leftarrow i[p] - 1$
 end

6 Conclusion

With the use of the component based resource allocation model in the cloud computing an efficient resource allocation strategy is described. This model will be helpful in the future resource allocation of the cloud and as more and more nodes will be added to the cloud, the information generated by the component based resource allocation will be very important. Further functionality can be added in any of the component to provide more enchantment.

References

- [1] Foster, I., Zhao, Y., Raicu, I., Lu, S.Y.: Cloud Computing and Grid Computing 360-degree compared. In: Proc. Grid Computing Environments Workshop, GeE 2008, pp. 1–10. IEEE Press (2008)
- [2] Peixoto, M., Santana, M., Estrella, J., Tavares, T., Kuehne, B., Santana, R.: A Metascheduler Architecture to provide QoS on the Cloud Computing. In: 17th International Conference on Telecommunications (2010)
- [3] Sun, D., Chang, G., Wang, C., Xiong, Y., Wang, X.: Efficient Nash Equilibrium Based Cloud Resource Allocation by Using a Continuous Double Auction. In: International Conference On Computer Design And Applications, ICCDA 2010 (2010)
- [4] Zhong, H., Tao, K., Zhang, X.: An Approach to Optimized Resource Scheduling Algorithm for Open-source Cloud Systems. In: The Fifth Annual China Grid Conference (2010)
- [5] Teng, F., Magoulès, F.: Resource Pricing and Equilibrium Allocation Policy in Cloud Computing. In: 10th IEEE International Conference on Computer and Information Technology, CIT 2010 (2010)
- [6] Wu, Z., Ni, Z., Gu, L., Liu, X.: A Revised Discrete Particle Swarm Optimization for Cloud Workflow Scheduling. In: International Conference on Computational Intelligence and Security (2010)
- [7] Foster, I., Kesselman, C., Lee, C., Lindell, R., Nahrstedt, K., Roy, A.: A Distributed Resource Management Architecture that Supports Advance Reservations and Co-Allocation. In: Intl. Workshop on Quality of Service (1999)
- [8] Chervenak, A., Foster, I., Kesselman, C., Salisbury, C., Tuecke, S.: The Data Grid: Towards an Architecture for the Distributed Management and Analysis of Large Scientific Datasets. *Jrnl. of Network and Computer Applications* 23, 187–200 (2001)
- [9] Raicu, I., Zhao, Y., Dumitrescu, C., Foster, I., Wilde, M.: Falcon: a Fast and Light-weight task Execution framework. In: IEEE/ACM Super Computing 2007 (2007)
- [10] Raicu, I., Zhao, Y., Foster, I., Szalay, A.: Accelerating Largescale Data Exploration through Data Diffusion. In: International Workshop on Data-Aware Distributed Computing (2008)
- [11] Buyya, R., Abramson, D., Giddy, J.: Nimrod/G: An Architecture for a Resource Management and Scheduling System in a Global Computational Grid. In: IEEE Int. Conf. on High Performance Computing in Asia-Pacific Region, HPC ASIA (2000)

Interval Evidential Reasoning Algorithm for Requirements Prioritization

Persis Voola¹ and A. Vinaya Babu²

¹ Department of Computer Science, Adikavi Nannaya University,
Rajahmundry, India
persisvoola@yahoo.co.in

² Department of Computer Science, JNT University, Hyderabad, India
dravinayababu@jntuh.ac.in

Abstract. A good requirements prioritization technique is one which involves all the relevant stakeholders, provides them the flexibility of assessing a requirement by means of subjective and uncertain inputs, and aggregates these assessments to produce reliable requirements priorities. This paper addresses this by applying Interval Evidential Reasoning (IER) algorithm. Analytic Hierarchy Process (AHP) is employed to determine the varying contribution of stakeholders. The degree of satisfaction with the requirements priorities will be obtained by following the same procedure followed for requirements assessment and aggregation.

1 Introduction

Requirements prioritization is an integral part of requirements engineering phase and this activity is significant for several reasons [1]. One such is to select a subset of the requirements and still meet the competing and conflicting needs and expectations of the various stakeholders. Requirements prioritization enables to identify a stable set of requirements to be implemented in the current release. Prioritization helps the Project Manager resolve conflicts, plan for staged deliveries and make the necessary trade-off decisions. A good number of prioritization techniques are available in the literature and some adopted by the industry as per their specifics. All of them assess a requirement using a single number on one of the ordinal, nominal or ratio scales and no single prioritization technique exists for intervals [2].

The following features were found to be *not present* with the current requirements prioritization techniques.

1. A provision to take into account the concern of multiple stakeholders by their value to the product.
2. A provision for expressing each requirement assessment with interval inputs. E.g.; Req1 is assessed to be of (grade1-grade3) where grade1 may be designated as low importance and grade3 of high importance.
3. A provision for expressing distribution of probabilities with inputs. E.g.; Req1 is assessed to be of (grade1-grade3) with 80% certainty and grade4 with 20%certainty.
4. A provision to accommodate ignorance in assessment. E.g.; Req 1 is assessed to be of unknown importance. Req1 is assessed to be of grade1 with 80% certainty and remaining 20% is unknown.

5. A provision for aggregating multiple stakeholders assessments who exert varying degrees of contribution in a consistent way.

6. A provision to generate combined degrees of belief like Req1 is assessed to grade1 with x% and grade2 with y% which is an aggregation of all the assessments of stakeholders.

The motivation for this paper is to introduce a requirements prioritization approach incorporating all the above features which are not present in any one of the existing prioritization techniques. The core idea is as follows:

“Identify the relevant stakeholders and place them in one of the critical, marginal, negligible categories as of their value to the organization. Their weights will be determined by Analytic Hierarchy Process (AHP). Allow the stakeholders to carry out the assessment with grade intervals and associated degrees of belief. Aggregate the assessments applying the Interval Evidential Reasoning algorithm to generate prioritized list of requirements. Obtain the degree of satisfaction with the prioritized list by following the same procedure for assessment and aggregation of a requirement value.”

Section 2 discusses determining the stakeholder weight applying AHP. Section 3 discusses about assessment of a requirement from the value perspective using grade Intervals and associated degrees of belief. Aggregation of assessment inputs from various stakeholders is addressed by Interval Evidential Reasoning algorithm and discussed in section 4. Section 5 discusses obtaining consensus with the prioritized list. We conclude with a discussion on future work in Section 6.

2 Stakeholder Weight

A Requirement Engineering activity begins with the relevant stakeholder identification and a good Requirements Prioritization technique is one which involves all the relevant stakeholders in the prioritization activity. The stakeholder concept was first introduced in a 1963 internal memorandum at the Stanford Research Institute. It defines stakeholders as “those groups with out whose support the organization would cease to exist” [8]. A stakeholder is defined as “person or organization who influences a system’s requirements or who is impacted by that system” [3]. IEEE-1471 defines as “An individual, team or organization with interests or concerns relative to a system”. They encompass business manager, project manager, marketing representatives, developers, end users, project sponsor or client or customer, architect, tester, quality engineer, product manager, operator and maintainer each with their own perspective of the product. Once they are identified, it is important to find the influence they exert on the project. The stakeholders influence to be managed in relation to the requirements to ensure a successful project. Their stake in the project is of varying levels and the problem is to assign weights to the stakeholders by the category. One sort of categorizing as in [4] is as Critical-if neglect might kill the project or render the system useless, Major – if neglect would have a significant negative impact on the system and Minor-if neglect would have marginal impact on the system. Mendelow’s power interest grid [5] categorizes stakeholders as High power-High Interest, High Power-Low Interest, Low Power-High Interest and Low Power-Low Interest. In [6]

stakeholders are grouped as exerting high, medium and low impact on the system. Several other means of categorizations are available [7]. This paper organizes the identified relevant stakeholders to be in one of the Critical, Marginal, and Negligible categories. Analytic Hierarchy Process is employed to determine the weighted impact of the stakeholders in each category and the Requirements Engineer or some other key stakeholders with knowledge on AHP will carry out this task.

2.1 Analytic Hierarchy Process

Thomas Saaty introduced a technique for multi criteria decision making called AHP[9]. It is a paired wise engine that generates relative measurement for the object for the decisions. The process of applying AHP with an example can be found in [15].

By applying the AHP to generate weights for the 3 stakeholder groups namely Critical (C), Marginal (M) and Negligible (N), the pair wise comparisons are as shown in Table 1.

Table 1. Pairwise Comparisons of Stakeholder groups

Stakeholder Group	Critical	Marginal	Negligible
Critical	1	3	6
Marginal	1/3	1	2
Negligible	1/6	1/2	1

The priority vector generated by applying AHP for this table is

$$[C \ M \ N] = [0.67 \ 0.22 \ 0.11]$$

It is assumed that all stakeholders in a category exert the same influence and hence the same weight to all the stakeholders in a category. Critical, Marginal, Negligible stakeholders' contribution in the requirements prioritization is 0.67, 0.22, 0.11 respectively of the total weight. Consistency Ratio = 0. This says that the pair wise comparisons are perfectly consistent.

3 Assessment of Requirements

This section discusses how and justifies why a requirement will be assessed using grade intervals and associated degrees of belief. Assessment can be precise or imprecise. But assessment only using precise answers always has the probability of being incorrect [12]. In many decision situations using a single number to represent a judgment proves to be difficult and sometimes unacceptable. Information would have been lost or distorted in the process of pre aggregating different types of information such as a subjective judgment, a probability distribution, or an incomplete piece of information into a single number [13] Intervals are necessary to describe degrees of belief [14] because even an expert can not assess a requirement to a precise number. Assessments inherent characteristic is uncertainty which can best be expressed using interval values.

Assessment to be done based on one or more of the several criteria. The criterion for prioritization in [15] is along the dimensions of value and cost. Wiegers[16] proposed an approach for prioritizing requirements along the dimensions of risk, penalty and cost. Babok [17] argues that during prioritization several criteria like value, cost, risk, difficulty of implementation, likelihood of success etc to be taken into account.

The criterion considered for assessment of a requirement is the *requirement value* i.e; the value a requirement provides to the stakeholders as this is the only criterion which is the focal point of all the stakeholders [10]. The value perspective is different for different stakeholders like increasing sales, increasing profit, finding new customers, beating competitors etc.[11] discusses the concern of how a requirement with high value can be neglected for the reason of other criteria like cost, risk, schedule etc. After the requirements are assessed and prioritized on value criterion does it make sense to make acceptable trade off among other conflicting goals like personal preference, business value, cost, schedule, effort, risk, requirements stability, legal mandate etc and it is the responsibility of the project manager and his team. All the stakeholders must be educated on the value dimension before assessment. Now the assessment activity will be carried out which incorporates grade intervals, uncertainty and ignorance in degrees of belief as these are integral parts of any assessment.

The set of evaluation grades considered for requirements assessment is as below:

$$G = \{L, A, H, U\}$$

$$= \{\text{Low importance, Average importance, High importance, Urgent importance}\}$$

The cardinality of the set must be small enough so not to impose over burden on the users and must be rich enough which covers all possibilities of assessment. These grades are sufficient as they cover all types of assessment starting from less important to urgently important. The assessment of a requirement can be a distribution of the grades in the set $\{L, A, H, U, L-A, L-H, L-U, A-H, A-U, H-U\}$

An illustration of assessment inputs is shown in table 2 for the requirements R1, R2 of a system and four stakeholders in each category assumed.

Table 2. Assessments by Stakeholders

Stakeholder Category	Stakeholder Identity	R1	R2
Critical Stakeholders (0.67)	C1	(A,1)	(H,1)
	C2	(H,0.7)(A,0.3)	(A-U,1)
	C3	(A,1)	(H,0.5)(A-U,0.5)
	C4	(H,0.8)(A,0.2)	(A-U,1)
Marginal Stakeholders (0.22)	M1	(L,1)	(H,1)
	M2	(L,0.8)(A,0.2)	(A-U,1)
	M3	(A,1)	(H,0.5)(A-U,0.5)
	M4	(L,0.9)(A,0.1)	(H,0.5)(A-U,0.5)
Negligible Stakeholders (0.11)	N1	(A,1)	(A-U,1)
	N2	(A,0.5)(L,0.5)	(A-U,1)
	N3	(A,1)	(A-U,1)
	N4	(A,1)	(H,1)

4 Interval Evidential Reasoning Algorithm for Aggregation of Assessments

Interval Evidential Reasoning (IER) is an extension of the Evidential Reasoning(ER) to accommodate intervals in assessment. ER algorithm is developed for aggregating multiple attributes based on a belief decision matrix and the evidence combination rule of the Dempster-Shafer theory. Aggregation process of the ER algorithm is a special case of the IER algorithm. If intervals are not present in the input ER can be used for aggregation. ER and IER have their prominence in a number applications.[13, 18, 19, 20, 21, 22]. This paper discusses how IER can be used to compute relative importance of the stakeholder requirements. Rich information for analysis like combined belief degrees across requirements, across stakeholders also can be obtained both in textual and graphical forms. A complete description of the algorithm can be found in [13].

The assessment from each group of stakeholders in the previous section is consolidated to produce table 3.

Table 3. Consolidated assessments

Stakeholder category	R1	R2
Critical (0.67)	(A,0.625)(H,0.375)	(H,0.375)(A-U,0.625)
Marginal (0.22)	(L,0.675)(A,0.325)	(H,0.5)(A-U,0.5)
Negligible (0.11)	(L,0.125)(A,0.875)	(H,0.75)(A-U,0.25)

Sample aggregation procedure for R1 by the Critical (CR1) and R1 by the Marginal (MR1) groups of the matrix given below. Remaining elements will be aggregated following the same procedure. The problem is to aggregate (A, 0.625) (H, 0.375) with (L, 0.675) (A, 0.325). The procedure follows.

Step1: Compute the Basic Probability Masses.

$$CR1_{AA} = 0.4187 \quad CR1_{HH} = 0.2512$$

Remaining probability mass $CR1_G = 1 - 0.67 = 0.33$

$$MR1_{LL} = 0.1485 \quad MR1_{AA} = 0.0715$$

Remaining probability mass $MR1_G = 1 - 0.22 = 0.78$

Step 2: Find Combined Probability Masses. To generate these, table 4 to be generated from which the values will be used in subsequent calculations.

Table 4. Combined Probability masses of CR1 and MR1

Aggregate (CR1,MR1)	CR1 _{AA} (0.4188)	CR1 _{HH} (0.2513)	CR1 _G (0.33)
MR1 _{LL} (0.1485)	0.0622 _φ	0.0373 _φ	0.0490 _{LL}
MR1 _{AA} (0.0715)	0.0027 _{AA}	0.0180 _φ	0.0235 _{AA}
MR1 _G (0.78)	0.3267 _{AA}	0.1960 _{HH}	0.2574 _G

The Combined Probability Mass (CPM) for each grade is generated by Summing all the Probability Mass elements (SPM) of that grade as given in table 4 and multiplying this with the scaling factor 1/(1-SEA). SEA is Sum of Empty Assessments.

$$CPM_{LL} = SPM_{LL}/(1-SEA) = 0.0555$$

$$CPM_{AA} = SPM_{AA}/(1-SEA) = 0.4307$$

$$CPM_{HH} = SPM_{HH}/(1-SEA) = 0.2220$$

$$CPM_G = SPM_G/(1-SEA) = 0.2917$$

Step3: Find Combined Belief Degrees. (CBD)

$$CBD_{LL} = CPM_{LL} / (1-CPM_G) = 0.0784$$

$$CBD_{AA} = CPM_{AA} / (1-CPM_G) = 0.6081$$

$$CBD_{HH} = CPM_{HH} / (1-CPM_G) = 0.3134$$

The aggregated assessment of critical and marginal stakeholders on R1 is (L,0.0784)(A,0.6081)(H,0.3134). The sum of combined belief degrees is 0.999. If there are no rounding errors, this sum will be 1. This after aggregating with the assessment of R1 by Negligible stakeholders (L,0.125)(A,0.875) is as follows. This process to be repeated for all assessments for each requirement and the final results are as shown below.

$$R1: (L, 0.0747)(A,0.6305)(H,0.2875)$$

$$R2: (A-U, 0.5382) (H, 0.4613)$$

Step4: Determine Ranks. It is always not clear by observing the aggregated assessments to determine the priority. For ranking requirements, expected utilities to be calculated. Because of interval uncertainties, the maximum, minimum and average expected values are calculated.

$$U_{max} = \sum_{i=L}^U \sum_{j=L}^U CBD_{ij} u(G_{ij}) \quad U_{min} = \sum_{i=L}^U \sum_{j=L}^U CBD_{ij} u(G_{ii}) \quad (1)$$

u(G) is the utility of the grade and is assumed to be equidistantly assigned like u(L) = 0.25, u(A) = 0.5, u(H) = 0.75, u(U) = 1

$$U_{avg} = (U_{max} + U_{min}) / 2. \quad (2)$$

Applying (1) and (2) for R1 and R2

$$R1: U_{min} = U_{max} = U_{avg} = 0.5496$$

$$R2: U_{min} = 0.6151 \quad U_{max} = 0.8842 \quad U_{avg} = 0.7497$$

These average values when arranged in sorted order give the ranking for the requirements and it is clear that R2 (0.7497) is of higher preference than R1 (0.5496). The above 4 step process is scalable to N number of requirements.

5 Obtaining Consensus

The prioritized list of requirements along with the supporting information like critical requirements that contribute to the success of the project, combined degrees of belief stakeholder group wise and requirements wise, percentage of preference of requirement over the other will be distributed to the stakeholders. They in turn will specify their degree of satisfaction with the sorted list in the form of grade intervals

and uncertainties as discussed in section 3. The evaluation grades can range from Low Satisfaction to Very High Satisfaction.

$$G = \{\text{Low, Medium, High, Very High}\}$$

These inputs will again undergo the aggregation procedure discussed in section 4. If the final outcome is nearer to ONE, it can be inferred that degree of satisfaction is high and sign-off to be obtained from all the stakeholders. Otherwise some form of consensus to be obtained with the dissatisfied stakeholders. If the list can not be materialized even after consensus discussion, redoing the assessment of requirements to be considered.

6 Discussion and Future Work

The strength of the approach lies in its novelty and rigor. This can be conveniently applied with the co located stakeholders or geographically distributed stakeholders. From the stakeholders' point of view, the approach introduced is friendly as it accommodates both quantitative and qualitative information, less time consuming, scalable and reliable. From the requirements engineer point of view aggregating the assessments applying the IER algorithm provides reliable results but this is complex process as it aggregates only 2 assessments at a time and this has to be applied iteratively until all assessments are aggregated.

Conventional methods like Multiple Attribute Utility Theory MAUT generate the same results as IER. But they don't generate rich information for analysis like combined belief degrees, percentage of preference of one requirement over the other. So, a decision to be taken in advance regarding which approach to follow MAUT without complementary information or IER with complex aggregation process.

The viability of this approach for non functional requirements i.e. Quality attributes to be explored. As IER includes complex aggregation process, a tool to be developed to generate the ranks to the requirements quickly and efficiently.

References

1. Donald, G.: Fire Smith Prioritising Requirements. *Journal of Object Technology* 3(8) (September-October 2004)
2. Hermann, A., Daneva: Requirements Prioritization Based on Benefit and Cost Prediction: An agenda for Future Research. In: *Proc. of the Intl. Conf. Requirements Engineering (RE 2008)*, pp. 125–134 (2008)
3. Glinz, M.: Stakeholders in Requirements Engineering. *IEEE Software* 28(1), 18–20, ISSN 0740-7459
4. Damian, D.: Stakeholders in Global Requirements Engineering: Lessons Learned from Practice. *IEEE Software* 24(2), 21–27 (2007), doi:10.1109/MS.2007.55
5. Mendelow, A.: Stakeholder Mapping. In: *Proceedings of the Second International Conference on Information Systems*, Cambridge, MA (1991)
6. Mc Geeu, R.A.: Stakeholder Identification and Quality Attribute Prioritisation for a Global Vehicle Control System. In: *Proc. of the Fourth European Conference on Software Architecture*, ISBN/ISSN: 978-1-4503-0179-4

7. Stakeholder Analysis,
http://en.wikipedia.org/wiki/Stakeholder_analysis
8. Freeman, R.E., Reed, D.L.: Stockholders and stakeholders: A new perspective on Corporate Governance. *California Management Review* 25(3), 88–106 (1983)
9. Saaty, T.: *The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation*. McGraw-Hill (1980)
10. Voola, P., Vinaya Babu, A.: Quality Case and A Simplified Approach to Quality Attributes Prioritization. In: *Proc. of the International Conference on Frontiers of Computer Science.*, IISc, Bangalore, India, ISBN: 978-81-921929-0-1
11. Regnell, B., Host, M., et al.: An Industrial Case Study on Distributed Prioritisation in Market Driven Requirements Engineering for Packaged Software. *Requirements Eng.* 6(1), 51–62 (2001) ISSN:09473602
12. Boehm, B.W., Fairley, R.E.: Software Estimation Perspectives. *IEEE Software*, 22–26 (November/December)
13. Xu, D.-L., Yang, J.-B., Wang, Y.-M.: The evidential reasoning approach for multi-attribute decision analysis under interval uncertainty. *European Journal of Operational Research* 174, 1914–1943 (2006), doi:10.1016/j.ejor.2005.02.064
14. Nguyen, H.T., Kreinovich, V., Zuo, Q.: Interval Valued Degrees of Belief: Applications of Interval Computations to Expert Systems and Intelligent Control. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 5(3), 317–358 (1997)
15. Karlsson, J., Ryan, K.: A Cost Value Approach for Prioritising Requirements. *IEEE Software* 14(5), 67–74
16. Wiegers, K.: First Things First: Prioritizing Requirements. *Software Development* 7(9) (September 1999)
17. <http://www.theiba.org/AM/Template.cfm?Section=Body>
of Knowledge BABOK Guide 2.0
18. Xu, D.-L., Yang, J.-B.: Intelligent Decision System for Self –Assessment. *Journal of Multi Criteria Decision Analysis* 12, 43–60 (2003), doi:10.1002/media.343
19. Yang, J., Xu, D.L.: Multiple Criteria Decision Analysis Applied to Safety and Cost Synthesis. *Journal of UK Safety and Reliability Society* 21(2) ISSN: 0961-7353
20. Huynh, V.-N.: Multiple Attribute Decision Making Under Uncertainty: The Evidential Reasoning Approach Revisited. *IEEE Transactions on Systems Man and Cybernetics*, <http://hdl.handle.net/2115/14531>
21. Chin, K.-S., Yang, J.-B., et al.: An Evidential Reasoning Interval Based Method for New Product Design Assessment. *IEEE Transactions on Engineering Management* 56(1), doi:10.1109/TEM.2008.2009792
22. Wang, Y.-M., Yang, J.-B., et al.: Environmental Impact Assessment using the Evidential Reasoning Approach. *European Journal of Operational Research* 174 (2006), doi:10.1016/j.ejor.2004.09.059

Evolutionary Based Secured Coding Technique for Mobile Communication Networks

Y.V. Srinivasa Murthy¹, Suresh Chandra Satapathy¹,
A.A.S. Saranya¹, and K. Sundeep Saradhi²

¹ Department of CSE,
Anil Neerukonda Institute of Technology & Sciences,
Visakhapatnam, Andhra Pradesh, India – 531 162
{urvishnu, sureshsatapathy, anu.saran91}@gmail.com

² Department of CSE,
D M S S V H Engineering College,
Krishna District Andhra Pradesh
sundeepsaradhi@gmail.com

Abstract. Arithmetic Coding is the way to encode the text to restrict unauthorized reading access. This paper is mainly concerned with providing security for messages in cellular networks. Encryption is very essential methodology in these days to keep our information in secure way. This paper clearly explains how to keep the information in secured way using evolutionary technology and a special encoding technique. Encryption of data traffic in cellular network is essential since it is vulnerable to eavesdropping. This project focuses on encrypting the data sent between the mobile stations and base stations using a stream cipher method. However, the keys for encryption are generated using an evolutionary computation approach termed Genetic Algorithm.

1 Introduction

In mobile communications the SMS has become popular means of communication by individuals and businesses. Also people sometimes exchange confidential information such as passwords or sensitive data. But in cellular networks these are vulnerable to the eavesdroppers and hackers. So there is a necessity to encrypt the message in order to safeguard the information. Encryption of data traffic in cellular network is essential since it is vulnerable to eavesdropping. This paper focuses on encrypting the data sent between the mobile stations and base stations using a stream cipher method. **Stream ciphers** operate on a bit-by-bit basis, producing a single encrypted bit for a single plaintext bit. Stream ciphers are commonly implemented as the exclusive-or (XOR) of the data stream with the key stream. The security of a stream cipher is determined by the properties of the key stream. A completely random key stream would effectively implement an unbreakable one-time pad encryption, and a deterministic key stream with a short period would provide very little security.

1.1 Vernam Cipher

In **Vernam cipher** the message is represented as a binary string (a sequence of 0's and 1's using a coding mechanism such as ASCII coding. The key is a truly random sequence of 0's and 1's of the same length as the message. The encryption is done by adding the key to the message exclusive or modulo 2, bit by bit. This process is often called as exclusive or and is denoted by XOR. The symbol used is \oplus .

1.2 Arithmetic Coding

Arithmetic coding is a form of variable-length entropy encoding used in lossless data compression. It provides an ideal length for each coded word and it assigns interval to a particular character based on the frequency and probability of that particular character in the given text. Arithmetic coding encodes the entire message into a single number, a fraction n where $(0.0 \leq n < 1.0)$. The entire set of data is represented by a rational number, which is always placed within the interval of each symbol. With data being added the number of significant digits rises continuously.

1.3 Cryptography

Cryptography can be defined as the conversion of data into a scrambled code that can be deciphered and sent across a public or private network.[3] The three types of algorithms are: Secret Key Cryptography (SKC): Uses a single key for both encryption and decryption, Public Key Cryptography (PKC): Uses one key for encryption and another for decryption, Hash Functions: Uses a mathematical transformation to irreversibly "encrypt" information.

1.4 Genetic Algorithm

A **genetic algorithm (GA)** is a search heuristic that mimics the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and search problems.[1] Genetic Algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions of optimization problems using techniques inspired by natural evolution such as inheritance, mutation, selection and cross over.[2]

The mobile user sends the message and we need to encrypt that message. We generate number of keys randomly by using Random Key Generation algorithm. We calculate the fitness function for the generated keys and sort them in the ascending order of the fitness values. We select specified number of keys which has maximum fitness values. Using Evolutionary technique we apply crossover to the selected keys and obtain new generation of keys. Repeat the process until number of iterations specified. As a result of last iteration we get the best key among generated keys.

Next step is to encode the received plain text and the key generated in the previous step by using Huffman encoding technique. In this technique, we calculate the frequency of each character and then calculate the probability of each character. Sort

the characters in ascending order of probabilities and construct the tree. By obtained tree we get the binary code for each character. Now we replace each character in the plain text with the binary values obtained in the Huffman encoding technique.

In the final step we encrypt the message using a stream cipher called vernam cipher. In vernam technique we apply XOR operation between the plain text and key. Finally we get the cipher text for the given plain text.

2 Cellular Networks

Cellular communication is seeing an explosive growth due to increased usage. However, it is vulnerable to eavesdropping which poses a threat to security and privacy of the user. It is therefore essential that the data traffic across the cellular communication network is encrypted. [5] A Cellular network consists of mobile stations attached to a base station (BS). A cluster of BS's which is fixed, is attached to a mobile telephone switching office which is connected to the public switched telephone network (PSTN). Cryptographic schemes are developed for protecting alphanumeric data since the emerging wired and wireless IP networks are vulnerable to eavesdropping. Thus in the case of the cellular network, the messages sent between the mobile station and the base station is encrypted using a stream cipher method.

The keys are generated randomly i.e. initial population and the new population is obtained by using Evolutionary technique known as Genetic algorithm.

2.1 GSM

GSM is a cellular network, which means that mobile phones connect to it by searching for cells in the immediate vicinity.

GSM differs from its predecessor technologies in that both signaling and speech channels are digital, and thus GSM is considered a *second generation* (2G) mobile phone system. This also facilitates the wide-spread implementation of data communication applications into the system.

The GSM standard has been an advantage to both consumers, who may benefit from the ability to roam and switch carriers without replacing phones, and also to network operators, who can choose equipment from many GSM equipment vendors. GSM also pioneered low-cost implementation of the short message service (SMS), also called text messaging, which has since been supported on other mobile phone standards as well. The standard includes a worldwide emergency telephone number feature.

GSM provides enhanced features over older analog-based systems, which are summarized below:

Total Mobility

The subscriber has the advantage of a Pan-European system allowing him to communicate from everywhere and to be called in any area served by a GSM cellular

network using the same assigned telephone number, even outside his home location. The calling party does not need to be informed about the called person's location because the GSM networks are responsible for the location tasks. With his personal chip card he can use a telephone in a rental car, for example, even outside his home location. This mobility feature is preferred by many business people who constantly need to be in touch with their headquarters.

High Capacity and Optimal Spectrum Allocation

The former analog-based cellular networks had to combat capacity problems, particularly in metropolitan areas. Through a more efficient utilization of the assigned frequency bandwidth and smaller cell sizes, the GSM System is capable of serving a greater number of subscribers. The optimal use of the available spectrum is achieved through the application Frequency Division Multiple Access (FDMA), Time Division Multiple Access (TDMA), efficient half-rate and full-rate speech coding, and the Gaussian Minimum Shift Keying (GMSK) modulation scheme.

Security

The security methods standardized for the GSM System make it the most secure cellular telecommunications standard currently available. Although the confidentiality of a call and anonymity of the GSM subscriber is only guaranteed on the radio channel, this is a major step in achieving end-to-end security. The subscriber's anonymity is ensured through the use of temporary identification numbers. The confidentiality of the communication itself on the radio link is performed by the application of encryption algorithms and frequency hopping which could only be realized using digital systems and signaling.[3]

Services

The list of services available to GSM subscribers typically includes the following: voice communication, facsimile, voice mail, short message transmission, data transmission and supplemental services such as call forwarding.

3 Encryption Using GA

3.1 Existing System Using Ant Colony Key Generation:

The existing system provides security by encrypting the message which involves generation of key by using Ant Colony Key Generation algorithm.^[4] In Ant Colony Key Generation Algorithm, key is generated randomly and it will calculate energy value for that key. It is accepted if it is greater than the threshold value which is specified by the user otherwise generate another key. Here we are not comparing the strength of keys with one another. We are accepting the key if it is just greater than the threshold value.

3.2 Proposed System

Our proposed system overcomes the limitations in the existing system. We generate the best key using an evolutionary technique called genetic algorithm. In this technique, we generate a set of keys randomly and calculate fitness function for each key. Select the best keys which have maximum fitness value and apply crossover to generate a new set of keys. [2] Now calculate the fitness value for new keys generated and select best keys among them. Again apply crossover and generate new keys. Repeat this process for specified number of iterations. Then select the best key which has maximum fitness value. Here the generation of best key is done by selecting a key among a huge number of keys and we are comparing the fitness values of all keys. Then encryption process is done by using verner cipher.

4 Genetic Algorithm Based Key Generation

Genetic algorithms are one of the best ways to solve a problem for which little is known. They are a very general algorithm and so will work well in any search space.[2] All you need to know is what you need the solution to be able to do well, and a genetic algorithm will be able to create a high quality solution. Genetic algorithms use the principles of selection and evolution to produce several solutions to a given problem.

Genetic algorithms tend to thrive in an environment in which there is a very large set of candidate solutions and in which the search space is uneven and has many hills and valleys. True, genetic algorithms will do well in any environment, but they will be greatly outclassed by more situation specific algorithms in the simpler search spaces. Therefore you must keep in mind that genetic algorithms are not always the best choice. Sometimes they can take quite a while to run and are therefore not always feasible for real time use.[2] They are, however, one of the most powerful methods with which to (relatively) quickly create high quality solutions to a problem. Now, before we start, I'm going to provide you with some key terms so that this article makes sense.

4.1 Procedure

Step 1: Initially generate a set of keys depending on the population size randomly of the specified key size.

Step 2: Retrieve the message i.e. plain text to be encrypted from the GUI.

Step 3: Calculate the fitness function for each chromosome i.e. the key using the following function:

$$Energy_i = (C_{ij} \in P_i) / Count(key)$$

Where C_{ij} - Charater at i^{th} row and j^{th} column

P_i is the plain text at i^{th} row.

Step 4: Now sort the chromosomes according to the fitness function.

Step 5: Read the crossover probability (cp) from the GUI and perform crossover

Number of new chromosomes (N) =
 $(cp / 100) * \text{initial population size}$

Step 6: Generate new set of chromosomes by multipoint crossover, now we get N new chromosomes

Step 7: Calculate the fitness function for these new chromosomes and add them to the keys in previous iteration.

Step 8: Sort the chromosomes and select the best set of population for the next iteration.

Step 9: Repeat the steps from 3 to 8 till certain specified number of iterations.

Step 10: Choose the topmost key from the last iteration as the key to be used for encryption.

5 Arithmetic Encoding

The aim of the arithmetic coding (AC) is to define a method that provides code words with an ideal length. Like for every other entropy coder, it is required to know the probability for the appearance of the individual symbols. The AC assigns an interval to each symbol, whose size reflects the probability for the appearance of this symbol. The code word of a symbol is an arbitrary rational number, which belongs to the corresponding interval. As mentioned before Arithmetic code emphasizes of encoding entire text all at a time.

5.1 Steps in Arithmetic Coding

The arithmetic code for an alphabet (set of symbols) may be generated by obtaining the probabilities of each character which represents the occurrence of that character in the given text.

Here we take the plain text and the key generated from the genetic algorithm as the input:

Step 1: Range of the entire data set will be a single value between 0 and 1 and is considered to be a rational number.

Step 2: Each of the divided sub ranges determines a unique character.

Step 3: Count of sub intervals is equivalent to the count of unique characters that appear in the data set and size is based on the probability of the character in the given data.

Step 4: Based on the last sub interval value there will be an internal division in the range corresponding to each symbol of data.

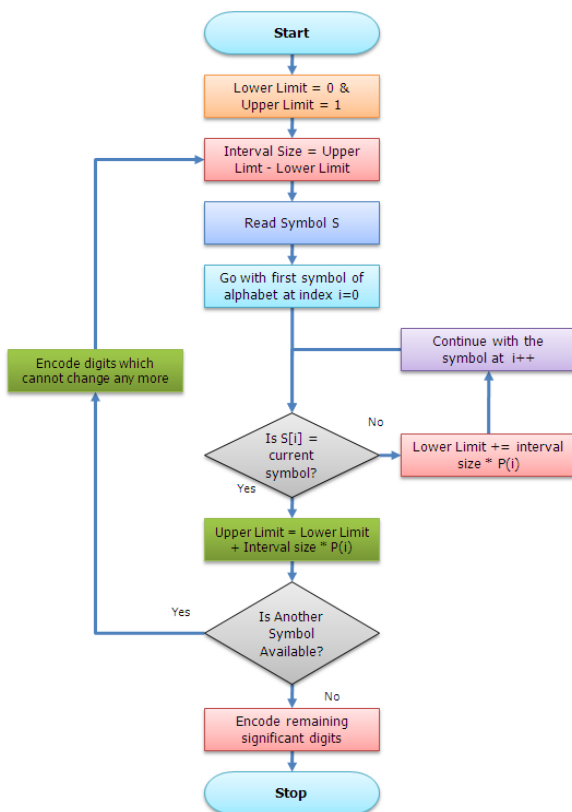


Fig. 1. Process of Arithmetic Coding

6 Vernam Cipher

The Vernam cipher is based on the technique of merging the plain text and the random text that is the key obtained from the GA. The resultant cipher obtained is free from all the risks that are ahead and can be sent in a safe from sender station. At the receiver's end the separation of key and plain text takes place thus retrieving the information that is sent.

Steps to implement vernam cipher:

Step1: By means of Using the encoding techniques like Arithmetic algorithm in this case generate an equivalent numerical value for the characters.

Step2: Encrypting key is generated by using GA technique.

Step3: XOR operation is performed between the numerical values of the characters of the plain text and the corresponding key value.

Step4: XOR operation on the receiver's side with the key generates the plain text again.

Step5: In order check the safety of the message XOR it with the key which generates the pad content again if it is unbiased.

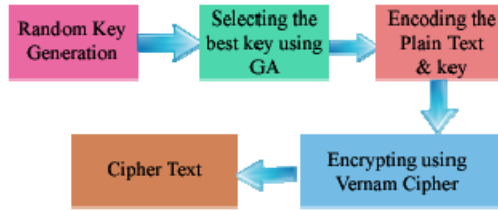


Fig. 2. Architecture of Key Generation

References

1. Davis, L.: Handbook of Genetic Algorithms. Von Nostrand Reinhold, New York (1991)
2. Goldberg, D.: Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, Redwood City (1988)
3. Pfleeger, C., Pfleeger, S.L.: Security in computing, 3rd edn. Prentice Hall of India (2003)
4. Wu, C.-P., Jay Kuo, C.C.: ³Design of Integrated Multimedia Compression and Encryption Systems. IEEE Transactions on Multimedia 7(5), 828–839 (2005)
5. Zhang, J., Stojmenovic, I.: Cellular Networks, University of Alabama, University of Ottawa, Canada
6. Fluhrer, S.R., Mantin, I., Shamir, A.: Weaknesses in the Key Scheduling Algorithm of RC4. In: Vaudenay, S., Youssef, A.M. (eds.) SAC 2001. LNCS, vol. 2259, pp. 1–24. Springer, Heidelberg (2001)

Research and Application of Dynamic Neural Network Based on Reinforcement Learning

Anil Kumar Yadav¹ and Ajay Kumar Sachan²

¹ Department of CSE,
IFTM University, Moradabad, U.P

² RITS, Bhopal
aky125@gmail.com,
sachanak_12@yahoo.com

Abstract. Dynamic neural network is became one of the most important approaches to eliminate Q table (look-up-table) of the machine intelligence. On the basis of comparison between general Artificial neural network and dynamic neural network, the development of dynamic neural network will be discussed. After the introduction of the theory and algorithms of reinforcement learning (RL), dynamic neural network will be applied as a basic decision taking unit (classifier neural network) in the form of a new technology. This will develop the application of reinforcement learning and provides a new idea for agent learning during real time operation. Use neural network for supervised learning, state as input/action as label. Reinforcement learning is widely use by different research field as intelligent control, robotics and neuroscience. It provides us possible solution within unknown environment. But at the same time we have to take care of its decision because RL can independently learn without prior knowledge or training and it take decision by learning experience through trail-and-error interaction with its environment.

In this paper, we discussed a new dynamic neural network model and its algorithms in detail, together with the issues that arise in Q table (look-up-table). Additionally, the benefit and challenges of reinforcement learning are described along with some of the problem domains where the dynamic neural network techniques have been applied. In order to access dynamic neural network is to eliminate Q table (look-up-table) and agent should learn during real time operation.

Keywords: Dynamic neural network, Machine learning, Reinforcement learning, Neural network classifier, Agent, State Action.

1 Introduction

With the development and prevalence of the machine learning, especially quarry based reinforcement learning. Learner (agent) during learning processes having a lot of training episodes. That will take large amounts time traditional general reinforcement learning technique like Q learning are facing the challenges in the field of such kinds of training episodes having more loop that affect the best decision path in original episodes. Dynamic neural network is introduced and it applied to predict

state as input/action label. Dynamic neural network is the extension of static neural network. The discussed a new models are developed based on static MLFL (multi layer feed forward neural network) .The architecture of artificial is neural networks is inspired by the biological nervous system [14]. It captures the information from data by learning and stores the information among its weights .this computation model is especially good, considering generalization and error to learner, compared to other computational models. In a departure from traditional neural networks, the structure and behavior of the dynamic model are closer to the nervous system. ANN has already been applied in transportation field. The most common network structure for short-term forecasting is MLFN which is a parallel distributed processing network. Parallel distributed processing network is formed by many basic units called neurons. Information processing takes place through the interactions of a large number of neurons can solve difficult tasks. This is the idea of learning tasks via different angles via neurons and the interactions between neurons.

In this research, we adopt MLFN as the basic model structure because of five major reasons. First, MLFN is a data-driven model and can learn the relationship between inputs and outputs from data without any assumption during model construction. This is beneficial because wrong model specification may lead to bad performance and it is a good idea to let data speak for itself. Second, MLFN is a nonlinear approximate which is appropriate for complex problems such as passenger behavior. In addition, MLFN is also a universal approximated, which means MLFN can approximate any kind of unknown functions. Third, MLFN is formed by many simple units and parallel connections. It is advantageous because information can be received by different units via connections, and distortion of some units would not cause too many damages. Forth, MLFN has been shown to have better performance than other parametric or non-parametric models in short-term traffic forecasting. It is a good option for model selection. Fifth, MLFN is flexible to extend to other kinds of neural network structures. MLFN has the most appealing properties so that many improvements and developments on this network structure are still going. Figure 1

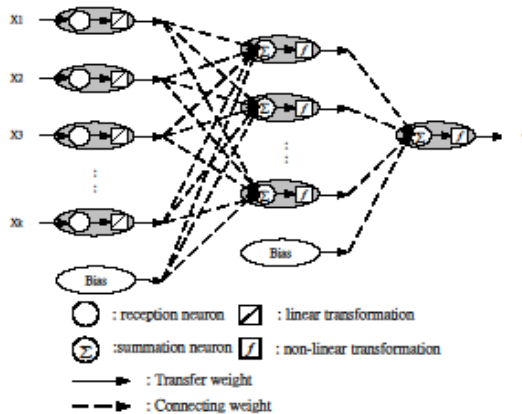


Fig. 1. MLFN Network structure

shows the whole network of MLFN applied in the study. The first layer is called input layer which just simply receives external information. The second layer is called hidden layer which can be seen as a feature extractor. If a network can capture the information of a task via the feature extractor as possible as it can, then it can be expected to have satisfactory performance. The third layer is called output layer which yields the final network output [14].

Reinforcement learning is an agent, which can perceive environment, learn to get the optimal action to achieve the target. When the agent makes any action, the environment will provide a feedback signal, which is called Reward. RL is learning what to do--how to map situations to actions--so as to maximize a numerical reward signal. In reinforcement learning, the learner [1] is a decision making agent that takes actions in an environment and receives reward for (or penalty) for its actions in trying to solve a problem. After of set of trial and error runs it should learn the best policy, which is the sequence of actions that maximize the total reward. An environment is represented by a set of states which an agent can perceive and take actions to change. Learning," allows the agent to operate in initially unknown environments and to become more competent than its initial knowledge alone might allow". Reinforcement learning refers to a collection of learning algorithms that seek to approximate solutions to stochastic sequential decision tasks with scalar evaluative feedback [11][15]. The computer then learns how to achieve that goal by trial-and-error interactions with its environment i.e. RL is defined as "learning what to do--how to map situations to actions so as to maximize a numerical reward signal. Fig.2 [4].

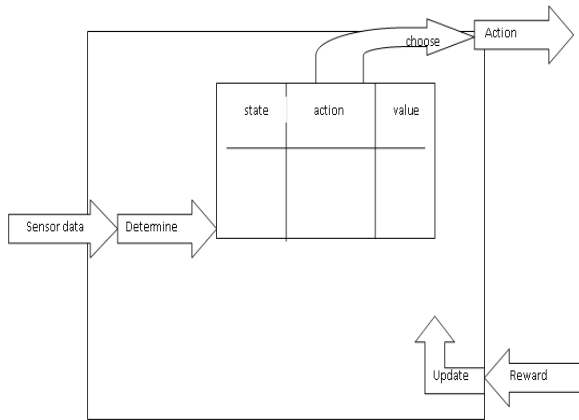


Fig. 2. The structure of Reinforcement Learning [4]

An agent interacts with an environment well defined by the fig.1.1 and this interaction yields an instantaneous scalar reward, serving as a measure of performance. The agent’s goal is to behave in a way that maximizes future reward.

We define an RL system [5] as a five tuple $\{S, A, \pi, RF, VF\}$, where S is a set of state of the environment, A is set of actions the agent can take, and the other elements are policy, reward function, and value function respectively.

A policy is a mapping from a state of the environment to an action to be taken by the agent. In other words, the policy dictates the behavior of the agent.

A reward function is a mapping from the state or state action pair of the environment to a numerical value called a reward signal. If the sequence of rewards received after step t is $r_{t+1}, r_{t+2}, r_{t+3}, \dots$, then the return from step t onward is: $R_t = r_t + \gamma V^{\pi}(s_{t+1})$. Where γ is a number ($0 < \gamma < 1$), called the discount rate. r_{t+1} is the reinforcement signal in $t+1$ moment. The purpose of the discount rate is twofold. The goal of an RL agent is to take actions to maximize expected return.

A value function of a given policy defines the expected return an RL agent can receive. The state-value function of a state s under a policy π is defined as $V^{\pi}(s) = E_{\pi}\{R_t | s_t = s\} = E_{\pi}\{\sum_{k=0}^{\infty} \gamma^k r_{t+k} | s_t = s\}$ this is the expected return under the policy π starting from state s . E is mathematical expectation [30][2]. A greedy method is one that always exploits. Exploration means taking an action other than the greedy action. The purpose of exploration is to discover other actions that might be better than the greedy action. The ϵ -greedy method is one that performs both exploitation and exploration. With probability $1-\epsilon$, where ϵ is a small positive number, it takes the greedy action, i. e., it exploits, and with probability ϵ it selects an action randomly. Any learning system basically contains 4 elements which are rewards, policy and environment: an environment is represented by a set of states. Learning agent: there is decision maker that perceives and select an action for the system. At present, evaluation of RL for control problem such as grid world using Temporal Difference network (TDN) techniques [3]. However, these methods address the problem of predicting time delayed rewards, they compute future rewards. Because they are value function as an estimate of future performance instead of sampled (commutative) rewards [6][9].

PROBLEMS WITH RL [3] In the years since the development of the machine learning technique especially in Query base self learning the learner (Agent) required a lot of training input of execution cycle. Now, we emphasize the another important problem associated with RL is that agent who travel in virtual world (called grid world), how to acquire knowledge efficiently by learning experience through trial-and-error interaction with its environment & behave intelligently and also aim is to reach the goal by moving shortest decision path randomly. While at each time (step) agents select randomly one of four actions: move Up, move Down, move Left, and move Right to perform, without communicating with each other.

In Recent time many reason works was done for RL and researchers has also proposed various algorithm and model such as SARSA [2][15], TDN [3]. Which tries to solve sequential decision making problem of continuous state and action space.

In this study, we focus on the model construction of artificial neural networks as dynamic neural network which acts as a decision making unit that eliminate Q table or look up table, in the real time operation.

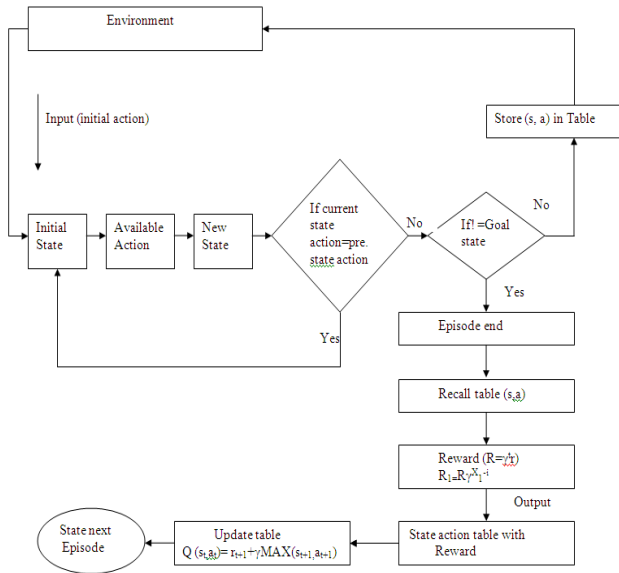
2 Problem

In the years since the development of machine learning technique especially in quarry based self learning the learner (agent) required a lot of training input of execution cycle [3].

Now we emphasize the another impatient problem associated with RL is that for making the best decision (or shortest path), the size of data increase drastically as the states and action of the system. In over come, so we required very large Q- table and very large memory to train the agent, So we are propose dynamic neural network to eliminate look-up-table with out much loss of information.

2.1 Previous Learning Agent Frame Work

We have already been investigate and evaluated learning agent frame work [12].



2.1.1 Working Mechanism

First, define the environment and state input; which we call state having the particular value s . Thus, the state action pair is presented to it in training. Next create a training module. First of all take random starting state and search available action take as an input from environment and achieve a new state. If previously taken state-action is equal to the current state then repeat and arrive starting state. Now examine goal state if goal achieved then update next episode. Otherwise, store state action pair (s_t, a_t) . again generate next state (s_{t+1}) using state-action pair. Then goal achieve.

Train the agent in the form of state-action pair. Then click on train, this lead to a new array editor window network. At this point you can view learned agent in the form of Q table. You can also see in the form of decision path. Now specify the inputs and goal. State-action pair (100×4) implemented for 10×10 grid world.

We had selected Q-learning (which has a characteristic of model free reinforcement learning, meaning we do not have any initial knowledge of the system, and we essentially try to find the optimal policy. It provides agent with the capability of learning to act optimally in Marko domains by experience action consequences function learning usually stores Q value relative with every state-action in a lookup

table).trail-and error run by state-action selection. Prior to training, state-action is initialized to random values. A training algorithm (Q-learning) is then used to progressively update by iteratively state action value. We had used Q-learning algorithm. Agent can be trained in the form of sate-action pair for function approximation (Q function). The training process requires state input and target output as a goal state. During training state-action value evaluated in terms of reward value, Q-learning algorithms use reward function to determine the state-action pair values using discount rate(γ).

2.2 Implemented Existing Q-Learning Algorithm

Q-learning is a form of model free reinforcement learning, meaning we do not have any initial knowledge of the system and we essentially try to find the optimal policy. Q-learning can be expressed as follows:

$$Q(s_t, a_t) = r_{t+1} + \gamma \max_a Q(s_{t+1}, a) \tag{1}$$

$$R_1 = R \gamma^{(x_1 - i)} \tag{2}$$

Where α is the learning rate and γ ($0 < \gamma < 1$) is discount rate. r_{t+1} is the reinforcement signal in $t+1$ moment. Essentially, the estimate for $Q(s_t, a_t)$, the value of the state-action pair at time t is updated using the best estimated value of the next state.

At any state action pair $Q(s_t, a_t)$ and single reward R_1 and next state possible. Where X_1 is the total steps that the agent move from initial state to the final reward state in episode. And i , is the count steps that the agent move from some initial state to the current state. γ is 0.9 discount rate. R_1 is the reward agent is given when it achieve the goal state and it is expressed as a numeric value (credit) with parameter γ , $X_{1, t}$. For each state action pair in the episode to learn more effective behavior (state action pair with large value).it is important that the policy is rational.

Reinforcement learning models have the advantage the learner is a decision making agent that takes actions in an environment and receives reward (or penalty) for its actions in trying to solve a complex problem based on set of trail-and-error runs; it should learn the best policy. Q learning usually stores Q value relative with every state action in a lookup table [35]. In Q learning there are sequences of episodes, learning steps repeat in every episode, in n^{th} time. In this work we have already been proposed and implemented algorithm [12] as fallows.

- Step 1 generate randomly starting state (s_n)
- Step 2 search available actions (a_n)
- Step 3 selects any one action randomly.
- Step 4 if checks previously taken same action then repeat from step one
- Step 5 now check for goal
- Step 6 if goal achieved than next episode
- Step 7 else, store (s_n, a_n) in temp array
- Step 8 update $Q_{n-1}(s_n, a_n)$ according to.

$$Q(s_n, a_n) = \begin{cases} (1 - \alpha)Q_n(s_n, a_n) + \alpha n[r_n + \gamma \max_a Q(s_{n+1}, a)] \\ r_n + \gamma \max_a Q(s_{n+1}, a_n + 1), \text{ if } \alpha n = 1 \\ Q(s_n, a_n) \text{ if } \alpha n = 0 \end{cases} \quad R_1 = R \gamma^{(x_1 - i)}$$

Step 9 now generate next state (s_{n+1}), using state action pair.

Step 10 if goal achieve then next
 Step 11 else repeat above step until stop criterion is satisfied

In Mat lab for train the agent in the form of state action pair (Q-table).In our implementation had size of (100x4) for <10x10> grid world.

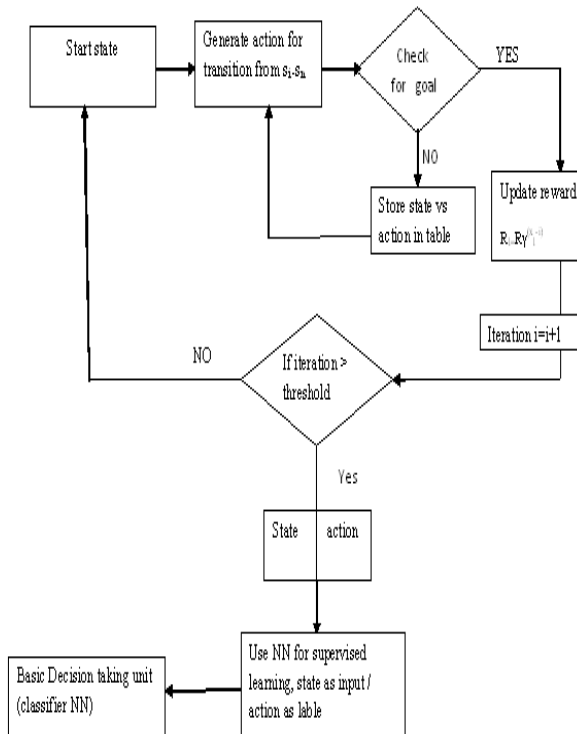
Where α is the learning rate and γ ($0 < \gamma < 1$) is the discount factor that reduces the influence of future expected rewards. So technically speaking, Q learning is evaluated Q value of each agent are evaluated by the sum of the rewards which the agent obtains during the previous one episode.

3 Proposed Model and Algorithm:

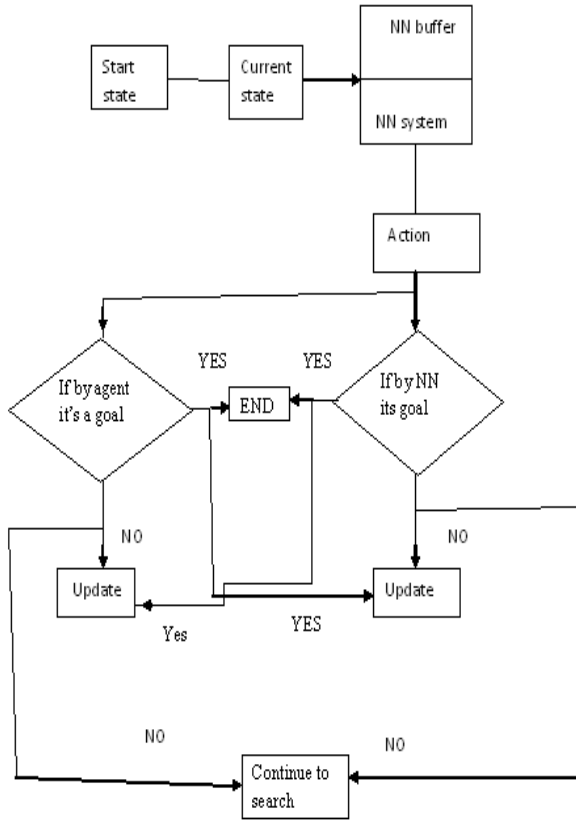
3.1 Dynamic Neural Network Training Model

In the study, we propose dynamic neural network training module as a (i) Agent on training (ii) agent on work.

Agent on Training:



Agent on work:



3.2 Dynamic Neural Network Algorithms

In this research, we are proposing new algorithms as follows.

(A) For basic agent training:

- Step: 1 generate random action a_n
- Step: 2 move to next state according to action
- Step: 3 if goal is not achieved
- Step: 4 go to step (1)
- Step: 5 update reward using $R_1 = R\gamma^{(x-1)}$
- Step: 6 check for iteration limit
- Step: 7 if under iteration limit go to step (1)
- Step: 8 else train NN by state action table or look-up-table or q table

(B) For Dynamic Neural network training

Step: 9 predict next state by NN

Step: 10 if decision by NN and agent both are same

(I) then predict next state is goal

(ii) Exit (achieved goal)

Step: 11 else (Both are not same)

Step: 12 Update NN

4 Discussion

In this research, we design dynamic neural network (DNN) model. We find that DNN is effective decision making unit (NN classifier) to take as an input/action as a label. Dynamic neural network eliminate q-table and agent should learning during real time operation. We have implemented, how to agent learn and to take random selection and learn short cut path from episode.

4.1 How to Agent Learn

In order to show how agent learn in the form of Q table for the grid world problem.

In order to evaluate the performance of machine learning techniques such as Query based reinforcement learning for the grid world problem, taken from figure 4.1. By observing an artificial agent who travels a virtual world called a grid world. In this grid world has the point for agent to start with, the point for the agent to aim as goal as a goal, usually we train “agent” who are to learn to behave intelligently, and agent’s aim is to reach the goal. We take the size of grid world is $M \times N$, where grid is square, M is equal to N as 10. the agent starts state position at the top left most. The goal is right most cells at the bottom, that is (10, 10). Agent can moves only one cell at a time to neighboring cell that is, up words, down words, to the right, or to the left, unless the agent touches the border or wall, if the action is possible. When the agent is touches the border or wall, the action that makes the agent cross the border is not performed but it must remain stooped or take decision for next available action. There are no rollback condition apply here. This would be repeated until the agent reaches the goal.

For the example, if the agent is at (1, 1) and the action is to down then agent remains at that point (2, 1), and if the next action is to right, then moves to (6, 4) or when the next action is down then agent moves to (8, 3). The grid world those agents are going to explore using a decision path is like a grid shown in above figure. Agent finds decision path from figure 3, the start position to goal position in 10×10 grid world using Q-learning proposed algorithm, agent select randomly one of four actions at a time.

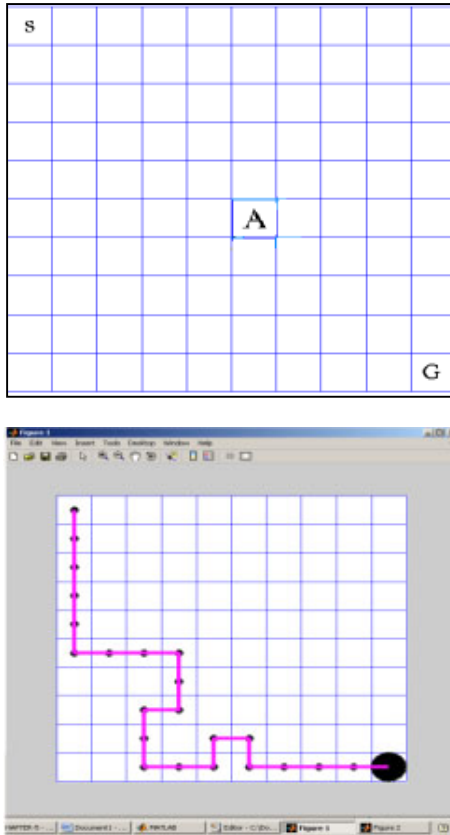


Fig. 3. In the Grid world of 10x10, the agent can move in four directions to find the goal

4.2 Random Selection and Shortcut Path

Agent moves one step upwards, downwards, to the right, or to the left, if the action is possible. If the movement is not possible due to the border of the grid world, do nothing and decide the next action again at random. This would be repeated until the agent reaches the goal. The maximum number of steps should be determined depending on the size of the grid world. We now look at the result of a random move by agent in a mentioned above. See the Figure. 4 Where an agent reaches the goal cell, it gains the reward 100.the value of discount rate parameter is set to be 0.9. Comprehensive way to remove loops and find shortcuts from episode for speeding up convergence, While the start cell is (2, 8) and fixed, the goal cell (9, 8) and is determined at random. The agent perceives its own coordinates (x, y), and has four possible actions to take: moving up, moving down, moving left and moving right, that is to say, it has actions to move into (x,y+1),(x,y-1),(x-1,y) and (x+1,y).some cell have walls at the boundaries with their adjacent cells, and the movement of the agent is blocked by the walls and the edge of the grid world. In addition, there are no roll back condition occurred here.

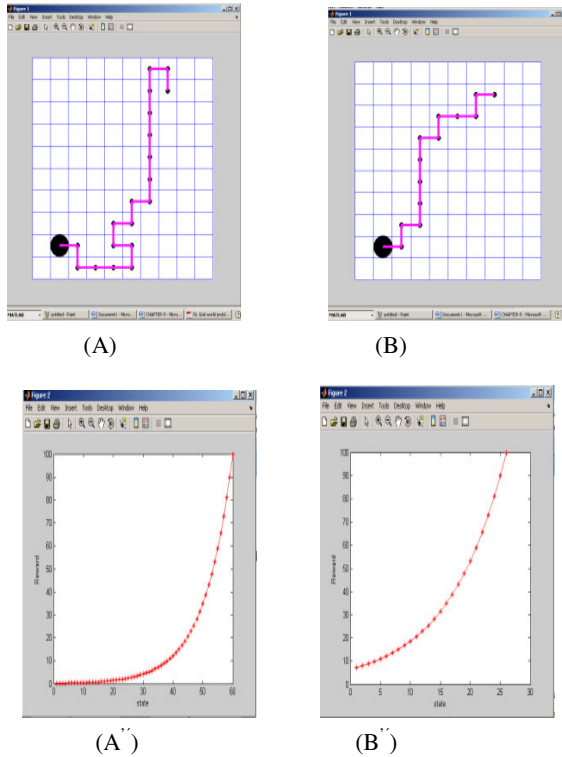


Fig. 4. In the grid-world of 10x10, starting from (2, 8) an Agent moves aiming the goal at (9, 2) of which the agent had no a-priori information. Left: A path chosen from 50 trials by random move to the goal and A' is the corresponding performance graph. Right: B route of the shortest path to the goal and B' is the corresponding performance graph. One trial had 200 episodes.

4.3 Evaluation Accuracy Assessments

Performance of the reinforcement learning algorithm using Q-learning measures can be used to assess Learning accuracy. E_Q is a measure of goal tracking efficiency and overall percentage of correctly evaluated training rate.

$$E_Q = \left[1 - \frac{\text{Minimum count step}(i) - \text{Total count step}(x_1)}{\text{Total program output}(T)} \right] \times 100$$

Where T is the total number of states, minimum count step is I and x_1 is total count step.

5 Conclusion and Future Work

In this study, we design dynamic neural network model to help agent in learning during real time operation. Dynamic neural network applies on quarry based reinforcement learning. Which eliminate the problem of look-up-table (Q table) or

large amount of training cycle? It should that knowledge acquired by the agent from environment and corresponding decision path.

Our proposed algorithm to learn shortest decision path in each and every episode. In this approach, look-up-table is removed to speed up convergence. While keep all state space knowledge acquired from original episode. To ensure each state in original episode can be refined by the reward acquired from achieving the final goal state and no any loss in state space knowledge.

Dynamic neural network is decision making system is known as learning agent. Reinforcement learning provides important mechanism for evaluation of machine learning problem such as control problem, robotics, weathering forecasting etc.

In the future work, reinforcement learning more effective by using combining multiple learner and other decision techniques like support vector machine and Genetic algorithm.

References

- [1] Alpayadin, E.: Introduction to machine learning. MIT press, Cambridge (2005)
- [2] Ima, H., Karo, Y.: Swarm Reinforcement Learning Algorithms Based on Sara Method, pp. 2045–2049. IEEE (2008)
- [3] Karbasian, H., Maida, N.: Improving Reinforcement Learning Using Temporal Deference Network EUROCON, pp. 1716–1722. IEEE (2009)
- [4] Quinn, L., Ming, C.Z.: The Research on the Spider of the Domain-Specific Search Engines Based on the Reinforcement Learning, pp. 588–592. IEEE (2009)
- [5] Trooper, J.W.C.: Optimizing Time Warp Simulation with Reinforcement Learning Techniques, pp. 577–584. IEEE (2007)
- [6] dam Silva, R.R., Claudio, A.: An Enhancement of Relational Reinforcement Learning, pp. 2055–2060. IEEE (2008)
- [7] Halmahera, K., Tadahiro: Effective integration of imitation learning and reinforcement learning by generating internal reward. In: Eighth International Conference on Intelligent Systems Design and Applications, pp. 121–126. IEEE (2008)
- [8] Taniguchi, T.: Role differentiation process by division of reward function in multi agent reinforcement learning, pp. 387–393. IEEE (2008)
- [9] Yang, Grace, D.: Cognitive Radio with Reinforcement Learning Applied to Heterogeneous Multicast Terrestrial Communication Systems, pp. 1–6. IEEE (2009)
- [10] Fang, Z., Tan, L.: Reinforcement Learning Based Dynamic Network Self-Optimization for Heterogeneous Networks, pp. 319–324. IEEE (2009)
- [11] Yadav, A.K.: Evaluation of Reinforcement Learning Techniques, pp. 1–4. ACM (2010)
- [12] Tsung-Hsien, Lee, C.-K.: Desin of dynamic neural network to forecast short-term railway passenger demand. Journal of the Eastern Asia Society for Transportation Studies 6, 1651–1666 (2005)
- [13] Li, H., Kozma, R.: A Dynamic neural network methed for time series prediction using the KIII model, pp. 347–352. IEEE (2003)
- [14] Lucian, Robert: A comprehensive survey of multiagent reinforcement learning, pp. 156–169. IEEE (2008)

An Effective Defence Mechanism for Detection of DDoS Attack on Application Layer Based on Hidden Markov Model

Suresh Limkar¹ and Rakesh Kumar Jha²

¹ Department of Computer Engineering,
GHRCEM, Pune, India

² Department of Electronics and Communication Engineering,
SVNIT Surat, India
{sureshlimkar, Jharakesh.45}@gmail.com

Abstract. There has been a lot of related work for detecting distributed denial of service attacks (DDoS) targeted on TCP and IP Layer. But these techniques cannot handle attacks which are mainly based on application layer. The severity of application layer DDoS attack has become a major threat to network operators nowadays. In this paper we introduce a new scheme based on hidden markov model (HMM) that distinguishes HTTP flooding attacks from legitimate HTTP traffic. An extended hidden Markov model is proposed to describe an anomaly. Each user's behavior is captured and profiled using HMM. In case of anomaly detection the user is authenticated using CAPTCHA otherwise added to the blacklist in case of attack, resulting in to blocking all the further requests from this user. We developed online e-banking portal to validate our Model. The experimental results show a distinctive and predictive pattern of the DDoS attacks, and our proposed model can successfully detect various DDoS attacks.

Keywords: DDoS, Hidden Markov Model, Anomaly detection, Captcha.

1 Introduction

As the complexity of Internet is scaled up, it is likely for the Internet resources to be exposed to Distributed Denial of Service (DDoS) attacks [1-3]. Distributed denials of service (DDoS) attacks constitute one of the major threats and are among the hardest security problem facing today's Internet [1]. Because of the seriousness of this problem, many defense mechanisms, based on statistical approaches [4-14], have been proposed to combat these attacks. Although the approaches based on statistics attributes are not always workable for some special DDoS attacks which work on the application layer. This has been witnessed on the Internet in 2004, when a worm virus named "Mydoom" [15] used pseudo HTTP requests to attack victim servers by simulating the behavior of browsers.

The challenge of detecting Application layer DDoS (App-DDoS) attacks is due to the following aspects. (i) The App-DDoS attacks utilize high layer protocols to pass through most of the current anomaly detection systems designed for low layers and

arrive at the victim web server. (ii) App-DDoS attacks usually depend on successful TCP connections, which make the general defense schemes based on detection of spoofed IP address useless. (iii) Flooding is not the unique way for the App-DDoS attack.

From the literature, few studies can be found that focus on the detection of App-DDoS attacks. Which are based on Data Mining [16], Neural Network [17], Markov chains [18] and etc. However, we cannot find many methods in the literature that emphasize the large-scale Web sites and their security at application. This paper proposes a new Probabilistic Hidden Markov Model (P-HMM) to describe the Web user behaviors and implement anomaly detection for flooding-based Application layer level DDoS attacks.

HMM which has widely applied in Speech Recognition, Character Recognition and DNA sequences clustering [9], is not widely used in the application of network security [11-13]. The main difference between HsMM and HMM is that the state duration is not a constant or exponentially distributed. [15] Has proved that the HMM is better than the HsMM in describing the unstable distribution and can describe the second order self-similarity and long-range dependence of network traffic which may change with the time. Because of these advantages, the HsMM can be used to describe Web user behaviors.

This paper is organized as follows. In Section 2, we describe the related works of our research. In Section 3, we present the assumptions made in our scheme and establish a new model for our scheme. In Section 4 we present experiment and result analysis. We conclude our work in Section 5.

2 Related Work

Most current DDoS-related studies focus on the IP layer or TCP layer instead of the application layer. [4] Devised a defense system called D-WARD installed in edge routers to monitor the asymmetry of two-way packet rates and to identify attacks. IP addresses [5] (which assume that attack traffic uses randomly spoofed addresses) and TTL values [6] were also used to detect the DDoS attacks. Statistical abnormalities of ICMP, UDP, and TCP packets were mapped to specific DDoS attacks based on the Management Information Base (MIB) [7]. [8] discovered the DDoS attacking signature by analyzing the TCP packet header against the well-defined rules and conditions. [9] Attempted to detect attacks by computing the ratio of TCP flags to TCP packets received at a web server and considering the relative proportion of arriving packets devoted to TCP, UDP, and ICMP traffic. [19] proposed techniques to extract features from connection attempts and classify attempts as suspicious or not. The rate of suspicious attempts over a day helped to expose stealthy DoS attacks, which attempt to maintain effectiveness while avoiding detection. Out of all these attacks [7] methodology extracts at least one valid attack precursor, with rates of false alarm of about 1%. Since the framework depends on MIB information alone, it is straightforward to use these Statistical signatures to implement MIB watches in common Network Management Systems.

3 Model Preliminaries and Assumption

The Hidden Markov model (HMM) has been successfully applied to a number of scientific and engineering problems [20]. Most studies found in the literature, however, implicitly assume that the duration of any system state is constant (i.e., a unit time in a discrete-time model) or exponentially distributed. This simplifying assumption is made because efficient computation algorithms have been well developed to deal with such HMMs. There are a few studies that discuss more general situations, where the duration of any state is explicitly assumed to be non exponential.

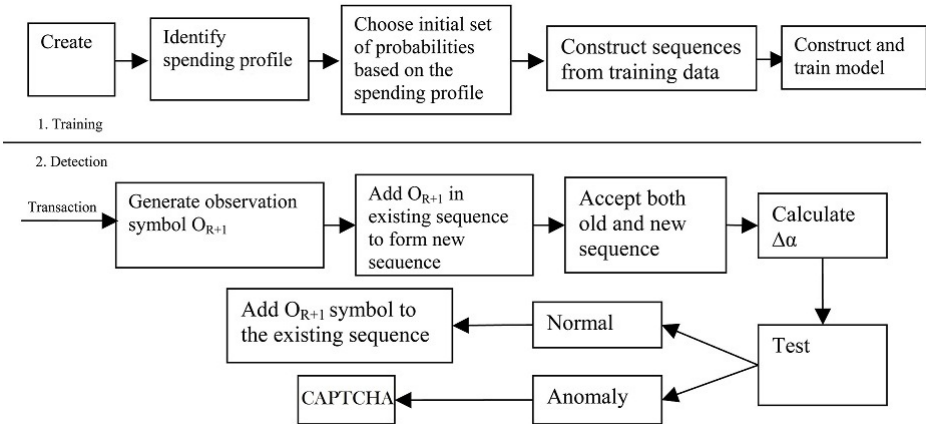


Fig. 1. Process flow of the proposed model

In our model given in fig.1 $(O_{R+1})^1$ we use Hidden Markov Models in order to extract web user behavior patterns from past one month behavior and subsequently compare an incoming request, and the latest effected request sequence of the same user, with the sets of extracted patterns. The comparison will generate a percentage value which represents the probability of the analyzed pattern being contained in that particular model. Based on the profiling mechanism applied previously we propose that independent HMMs are devised for each profile. Such an implementation will increase the effectiveness of HMMs since accessing patterns that are popular in one profile may not be popular in another profile and therefore the HMMs built will be specific for each group of users.

An HMM has a finite set of states governed by a set of transition probabilities. In a particular state, a outcome or observation can be generated according to an associated probability distribution. It is only the outcome and not the state that is visible to an external observer.

¹ (O_{R+1}) means adding new observation in existing observation database.

3.1 Modeling the User Behavior

To model the web user behavior we made following assumptions:

- Web user behavior can be profiled using the web page request sequence.
- It is very difficult for an attacker to build a routine that exactly mimics the legitimate user behavior. Because user browsing can be modeled by three aspects Web page request rate, web page reading time and the web page request sequence. An attacker cannot generate a request sequence which mimics legitimate user sequence since an attacker uses a predefined routine which cannot capture the dynamics of user and network.
- A http request sequence can be simulated in following ways
 1. Attacker's routine generates the http request objects randomly which makes the attackers request sequence to differ from normal users request sequence.
 2. The attacker has to embed a predefined set of http request list into the attacking routine, so it becomes very easy to identify that some users are accessing the same web pages periodically.
- Once the attacker gains the control of zombies (infected computers) attacker distributes a varying http request sequence list to each zombie periodically but still as attacker cannot view the users access patterns stored on victim server and it will be different from users behavior profile build on the victim computer and hence can be easily identified.

3.2 Behavior Tracking Algorithm

In recent years, [21] investigated the capabilities of HMM in anomaly detection. They classify TCP network traffic as an attack or normal using HMM. [22] Suggested an HMM-based intrusion detection system that improves the modeling time and performance by considering only the privilege transition flows based on the domain knowledge of attacks. [23] Proposed the application of HMM in detecting multistage network attacks. [24] Used HMM to model human behavior. Once human behavior is correctly modeled, any detected deviation is a cause for concern since an attacker is not expected to have a behavior similar to the genuine user. Hence, an alarm is raised in case of any deviation. Our behavior tracking algorithm is having following steps.

- 1 In this model we build two machines i) 50 % probability machine (which has 0.50% probability for OK, 0.50% probability for Fraud). ii) Truth machine (which has 0.95 % Probability for OK, 0.05% Probability for Fraud).
- 2 Sequence generation for truth machine: 200 rows are generated using markov generator algorithm .Each row contains 100 sequences.
- 3 Learning Machine: BaumwelchLearner Algorithm is used to create object of learning machine.
- 4 0.50% machine is initialized. The distance between 0.50% machine and Truth machine is calculated by using Kullback Leibler Distance Calculator algorithm.
- 5 Here to normalize the machine we take 10 iterations. The test sequence is Generated and inputted to learn the machine.

- 6 If the probability of test sequence is between 0 to 0.7 then it results OK, and if it is between 0.7 to 1 then it results FRAUD.
- 7 If the result of the test sequence is FRAUD server returns CAPTCHA to that client.
- 8 If server does not get response to the CAPTCHA for three consecutive replies and the difference between two consecutive request is less than some threshold value (here 20 msec) then server blocks all the requests from that ip address.

3.3 Anomaly Detection Module

If the resource is scarce and the server gets busy then we apply the filter for the incoming requests. Here a history is maintained for each client (IP) address which stores the latest 20 requests. If an unusual behavior is detected then server calls the behavior tracking module which reply's with the captcha to that client if server do not receive correct answer for the consecutive reply's of CAPTCHA then it checks the history of request sequence, now if the difference between two consecutive request is less than some threshold (50 msec in this case) then server blocks all the requests coming from that IP address and returns Block message.

Online eBanking portal was attacked by the application layer DDoS attack. The attack continuously called LoginPage.htm and Login_Error.htm. Where the later page does not exists there on the server and is just called to confuse the server. Initially server returned a login page for three times and upon getting no response since the attack did not respond to the reply, server returned three CAPTCHA pages and immediately blocked the IP address 192.168.1.1 from where the attack was made.

4 Experiment and Result Analysis

Testing end user request is termed as transaction. Using real data set is a difficult task. Banks do not, in general, agree to share their data with researchers. There is also no benchmark data set available for experimentation. We have, therefore, took a trained system. A trained data is used to generate a mix of genuine and fraudulent transactions. The number of fraudulent request in a given length of mixed request is normally distributed with a user specified μ (mean) and σ (standard deviation), taking end users spending behavior into account. μ specifies the average number of fraudulent request in a given request mix. In a typical scenario, an issuing bank, and hence, its DDoS receives a large number of genuine request sparingly intermixed with fraudulent request. The genuine requests are generated according to the end users most likely behavior. We have studied the effects of spending group and the percentage of request. We use standard metrics—True Positive (TP) and False Positive (FP), as well as TP-FP spread and Accuracy metrics, as proposed in [25] to measure the effectiveness of the system. TP request the fraction of fraudulent request correctly identified as fraudulent, whereas FP is the fraction of genuine request identified as fraudulent. Most of the design choices for a DDoS that result in higher values of TP, also cause FP to increase. To meaningfully capture the performance of

such a system, the difference between TP and FP, often called the TP-FP spread is used as a metric. Accuracy represents the fraction of total number of request (both genuine and fraudulent) that have been detected correctly. It can be expressed as follows:

“Accuracy is defined as sum of no of good request detected as good and no of bad request detected as bad divide by two”. We first carried out a set of experiments to determine the correct combination of HMM design parameters, namely, the number of states, the sequence length, and the threshold value. Once these parameters were decided, we performed comparative study with another DDoS.

4.1 Choice of Design Parameter

We considered the μ and σ , values to be 1.0 and 0.5, respectively. This is chosen so that, on the average, there will be 1 fraudulent request in any incoming sequence with some scope for variation. After the parameter values are fixed, we will see in

For parameter selection, the sequence length is varied from 5 to 25 in steps of 5. The threshold values considered are 30 percent, 50 percent, 70 percent, and 90 percent. The number of states is varied from 5 to 10 in steps of 1. We consider both TP and FP for deciding the optimum parameter values. An initial set of five simulation runs, each with 100 samples, was carried out to estimate the mean and standard deviation of both TP and FP for a fixed sequence length, number of states, and threshold value. Mean TP was found to be an order of magnitude higher than mean FP. Standard deviation of TP was 0.1 and that for FP was 0.005. We set the target 95 percent confidence interval (CI) for TP and FP, respectively, as ± 2.5 percent and ± 0.25 percent around their mean values. Using Student's t-distribution, the minimum number of simulation runs required for obtaining desired CI for TP. Since it is not convenient to present the detailed results for each of the 120 combinations. Thus, our design parameter setting is given as follows:

1. Number of hidden states $N = 10$,
2. Length of observation sequence $R = 15$,
3. Threshold value = 50%, and
4. Number of sequences for training = 100.

With this design parameter setting, we next proceed to study the performance of the system under various combinations of input data.

4.2 Performance Analysis

Based on the parameter in section 4.1, our system can therefore, correctly detect most of the transactions. However, when there is no profile information at all, the system shows some performance degradation in terms of TP-FP. This observation highlights the importance of trained data. Also, when there is little difference between genuine request and malicious request, most of the web server suffers from DDoS performance degradation, either due to a fall in the number of TPs or a rise in the number of FPs.

5 Conclusion

In this paper, we focused on early detection and prevention of DDoS attack using a new technique based on Hidden Markov Model. To test the model we designed a sample online banking portal. The server stores behavior characteristic of each user and in case if it finds unusual behavior or anomaly detection for a particular client then a CAPTCHA is returned to that client. For correct response server authenticates the client, whereas in case of attack the client ignores the CAPTCHA pages and makes more requests simultaneously (More than the threshold value) then server immediately blocks all the requests from this client and sends a BLOCK message for further requests.

For generation of DDoS we used a script in java which continuously requests Login_Page.htm and LOGIN_Error.htm. Once we run this script server returned three CAPTCHA Pages and then upon getting no response, server blocked this client immediately by sending a “Block” message.

References

- [1] Garber, L.: Denial-of-Service Attacks Rip the Internet. *IEEE Computer* 33(4), 12–17 (2000)
- [2] Houle, J.K., Weaver, M.G.: Trends in Denial of Service Attack Technology. CERT Coordination Center (2001)
- [3] Moore, D., Voelker, G.M., Savage, S.: Inferring Internet Denial-of-Service Activity. In: Proceedings of the 10th USENIX Symposium, pp. 9–22 (2001)
- [4] Mirkovic, J., Prier, G., Reiher, P.L.: Attacking DDoS at the source. In: Proc. 10th IEEE Int. Conf. Network Protocols, pp. 312–321 (September 2002)
- [5] Jin, C., Wang, H., Shin, K.G.: Hop-count filtering: An effective defense against spoofed traffic. In: Proc. ACM Conf. Computer and Communications Security, pp. 30–41 (2003)
- [6] Peng, T., Mohanarao, K.R., Leckie, C.: Protection from distributed denial of service attacks using history-based IP filtering. In: Proc. IEEE Int. Conf. Communications, vol. 1, pp. 482–486 (May 2003)
- [7] Cabrera, J.B.D., et al.: Proactive detection of distributed denial of service attacks using MIB traffic variables a feasibility study. In: Proc. IEEE/IFIP Int. Symp. Integrated Network Management, pp. 609–622 (May 2001)
- [8] Limwivatkul, L., Rungsawangr, A.: Distributed denial of service detection using TCP/IP header and traffic measurement analysis. In: Int. Symp. Communications and Information Technologies 2004 (ISCIT 2004), Sappom, Japan, October 29 (2004)
- [9] Noh, S., Lee, C., Choi, K., Jung, G.: Detecting Distributed Denial of Service (DDoS) Attacks Through Inductive Learnin. In: Liu, J., Cheung, Y.-M., Yin, H. (eds.) IDEAL 2003. LNCS, vol. 2690, pp. 286–295. Springer, Heidelberg (2003)
- [10] Ranjan, S., Swaminathan, R., Uysal, M., Knightly, E.: DDoS-resilient scheduling to counter application layer attacks under imperfect detection. In: Proc. IEEE INFOCOM (April 2006), <http://www-ece.rice.edu/~networks/papers/dos-sched.pdf>
- [11] Chang, R.K.C.: Defending against flooding-based distributed denial-of-service attacks: A tutorial. *IEEE Commun. Mag.*, 43–51 (October 2002)

- [12] Turchanyi, G., Mohacsi, J.: IPv4-IPv6 Transition- Just to cut the Gordian Knot? In: The 13th International Telecommunications Network Strategy and Planning Symposium (2000)
- [13] Gunderson, S.H.: Global IPv6 statistics- Measuring the Current State of IPv6 for ordinary Users, Google White Paper (2008)
- [14] Nagaraj, S., et al.: A Comparative Study of IPv6 Statistical Approach. IJCSE, International Journal on Computer Science and Engineering 02(04) (2010)
- [15] MyDoom virus,
<http://www.us-cert.gov/cas/techalerts/TA04-028A.html>
- [16] Florez, G., Bridges, S.A., Vaughn, R.B.: An Improved Algorithm for Fuzzy Data Mining for Intrusion Detection. In: Proceedings, 2002 Annual Meeting of the North American, Fuzzy Information Processing Society, NAFIPS, pp. 457–462 (2002)
- [17] Mukkamala, S., Janoski, G., Sung, A.: Intrusion Detection Using Neural Networks and Support Vector machines. In: Proceedings of the 2002 International Joint Conference on Neural Networks, IJCNN 2002, vol. 2, pp. 1702–1707 (2002)
- [18] Xing, D.-S., Shen, J.-Y.: A New Markov Model for Web Access Prediction. Computing in Science and Engineering 4(6), 34–39 (2002)
- [19] Basu, R., Cunningham, K.R., Webster, S.E., Lippmann, P.R.: Detecting low-profile probes and novel denial of service attacks. In: Proc.2001 IEEE Workshop on Information Assurance and Security, pp. 5–10 (June 2001)
- [20] Rabiner, L.R.: A tutorial on hidden Markov models and selected application in speech recognition. Proc. IEEE 77, 257–286 (1989)
- [21] Joshi, S.S., Phoha, V.V.: Investigating Hidden Markov Models Capabilities in Anomaly Detection. In: Proc. 43rd ACM Ann. Southeast Regional Conf., vol. 1, pp. 98–103 (2005)
- [22] Cho, S.B., Park, H.J.: Efficient Anomaly Detection by Modeling Privilege Flows Using Hidden Markov Model. Computer and Security 22(1), 45–55 (2003)
- [23] Ourston, D., Matzner, S., Stump, W., Hopkins, B.: Applications of Hidden Markov Models to Detecting Multi-Stage Network Attacks. In: Proc. 36th Ann. Hawaii Int'l Conf. System Sciences, vol. 9, pp. 334–344 (2003)
- [24] Lane, T.: Hidden Markov Models for Human/Computer Interface Modeling. In: Proc. Int'l Joint Conf. Artificial Intelligence, Workshop Learning about Users, pp. 35–44 (1999)
- [25] Stolfo, S.J., Fan, D.W., Lee, W., Prodromidis, A., Chan, P.K.: Cost-Based Modeling for Fraud and Intrusion Detection: Results from the JAM Project. In: Proc. DARPA Information Survivability Conf. and Exposition, vol. 2, pp. 130–144 (2000)

Author Index

- Abhaikumar, V. 887
Acharya, Kallol 145
Agrawal, Abhishek 879
Ahlawat, Savita 119
Ahmed, Mushtaq 103
Allam, Appa Rao 437
Anand, Sesham 1, 669
Anantharaman, Narayanan 755
Anil Kumar, Ravva 227
Anitha, T.N. 47
Aruldoss Albert Victoire, T. 291
Arumugam, G. 21
Arun, R. 531
Aruna Kumari, Devarakonda 379
Arunbhaskar, M. 861
Avadhani, P.S. 803, 819
- Bagchi, Kallol 59
Balajee, Maram 67
Bansal, Rohit 83
Barik, Alaka 319
Basu, Subhadip 111, 145, 639, 649, 739, 837
Benala, Tirimula Rao 75
Bhadoria, Robin Singh 83
Bharath, K. 47
Bhardwaj, Satyam 245
Bhateja, Vikrant 245, 779
Bhattacharya, Bidishna 93, 387
Bhattacharya, Prasanta 827
Bhavanasi, Venkat Ramana 591
Bhurani, Parvati 103
Biswas, Utpal 731
Boggavarapu, L.N. Phaneendra Kumar 353
- Bramaramba, Kakaraparathi 299
Brindha, G.R. 1
- Chaki, Rituparna 657
Chakrabarti, Amlan 613
Chakrabarty, Saikat 127
Chakraborty, Niladri 93
Chakravorty, N. 387
Chandra Bose, Divya 531
Chatterjee, Piyali 111
Chauhan, Abhishek Singh 83
Chintakayala, Prasad 807
Choudhary, Amit 119
Chowdhury, Soumit 209
- Damodaram, A. 591, 605
Darbar, Rajkumar 195
Das, Achintya 195
Das, Asit Kr. 507
Das, Asit Kumar 127, 137
Das, Nibaran 145, 739
Dash, S.S. 861
Dehuri, Satchidananda 75, 489
Devaraj, D. 515
Dheeba, J. 153
Dixit, Manish 83
- Elliot, Alex Christopher 179
Faruqi, M.A. 273
- Game, Pravin 763, 907
Geetamma, Tummalapalli 187
George, Teresa 715
Ghatol, Ashok A. 217
Ghosal, Soumya 195

- Ghosh, Shameek 755
 Ghoshal, Nabin 209
 Ghoshal, Sakthi Prasad 405
 Ghumbre, Shashikant U. 217
 Gopalakrishnan Nair, T.R. 461
 Govardhan, A. 669, 703
 Gupta, Manoj Kumar 237
 Gupta, Prateek 245
 Gupta, Rohit 779
 Gupta, Sumit 227
- Harikrishna Reddy, R. 169
 Harinie, Thiagarajan 887
 Hazra, Simanta 769
 Heymaraju, Ch. 845
- Jacob, Minu 11
 Jagadeeshkumar, M. 861
 Jagadish, Gurrala 345
 Jana, Nanda Dulal 281, 769
 Janani Chellam, Ilangovan 887
 Jasper, J. 291
 Jayachitra, V.P. 571
 Jayaraman, Valadi K. 755
 Jeya Nachiaban, N.M. 429
 Jha, Rakesh Kumar 943
 JoePrathap, P.M. 1
 Jyotishree 39
- Kaladhar, D.S.V.G.K. 161, 169
 Kaligathi, Katyayani 309
 Kar, Asutosh 319
 Kar, Rajib 405
 Kartheek, Vadlamani 853
 Khan, Habibulla 379
 Killani, Ramanji 337
 Kiran, S. Harish 861
 Kiran Kumar, R. 469
 Kishan Rao, K. 679
 Koteswara Rao, S. 299
 Kotha, Sita Kumari 413
 Koushik, S. 631
 Krishnaiah, Jallu 273
 Krishnamurthi, Ilango 583
 Krishna Prasad, A.V. 605
 Krishna Prasad, P.E.S.N. 469
 Krishnaveni, V. 21
 Kumar, C.S. 273
 Kumar, Rakesh 39
 Kundu, Mahantapas 111, 145, 739
- Lakshmi, P.V. 169
 Lakshmi, Sathya 11
 Lalitha Bhaskar, D. 803
 Latha, M. Madhavi 379
 Limkar, Suresh 943
- Madhu, Ramarakula 363
 Madhurakshara, S. 75
 Maheshwari, Divya 689
 Maheswari, Santhoshkumar 723
 Majumder, Koushik 371
 Mala, T. 547, 555
 Malakar, Samir 739
 Mallidi, Prudhvi Ravi Raja Reddy 437
 Manda, Kalyani 29
 Mandal, Durbadal 405
 Mandal, Jyotsna Kumar 209, 395, 623
 Mandal, Kamal K. 93, 387
 Mandal, Sangeeta 405
 Manne, Suneetha 413, 421
 Marjit, Ujjal 731
 Masilamani, Roberts 11
 Maulik, Ujjwal 837
 Meena, Yogesh Kumar 103
 Meyyappan, T. 429
 Misra, Arun K. 787
 Misra, Manoj 237
 Mohankrishna, Samantula 689
 Mudunuri, Suresh B. 437
 Mukhopadhyay, Somnath 59, 395
 Murthy, J.V.R. 265, 445
- Nagendra Kumar, Dirisala J. 265
 Nageswara Rao, D. 345
 Nageswara Rao, P.V. 169
 Naidu, Ch. Demudu 845
 Naik, Anima 453
 Narasimham, Challa 67
 Nasipuri, Mita 111, 145, 639, 649, 739
 Nath, Hiran V. 531
 Nath, Ravinder 319
 Navya, B. 353
 Neogi, Biswarup 195
 Nirmala Devi, Rangisetty 679
 Niyogi, Rajdeep 237
- Padhy, Isha 489
 Padhy, Sasmita 481
 Padmini, Sankaramurthy 715

- Pallamsetty, S. 437
 Panda, Ashok Kumar 489
 Panda, Sidhartha 481
 Patel, Omprakash 497
 Pati, Soumen Kr. 507
 Patnana, Sujana 437
 Perkins, A. Louise 179
 Petchinathan, G. 515
 Phadikar, Santanu 137
 Plewczynski, Dariusz 837
 Prabhakara Rao, G. 309
 Pradeep, Tummala 819
 Pradhan, Buddhadeb 523
 Prakash, S. 1
 Prasad, Bhanu 827
 Prasada Reddy, P.V.G.D. 445
 Prasad Reddy, P.V.G.D. 227, 845, 853
 Pratihar, Dilip Kumar 563
 Praveen, K. 531
 Preethi, J.D. 255
 Priscilla, R. 539
 Pullabhatla, Srikanth 807
 Pullela, S.V.V.S.R. Kumar 265
 Purohit, Lalit 497

 Raja Rajeswari, K. 299
 Rajasekhara Rao, K. 29
 Rajavel, Rajkumar 547, 555
 Rajendra, Rega 563
 Raju, S. 887
 Ramachandran, Sumalatha 571
 Ramachandran, Vivek Anandan 583
 Ramakrishna Murty, M. 445
 RamaPraba, Pazhayanoor Seethapathy 597
 Ramesh Kumar, Y. 67
 Ranganathan, H. 597
 Rashmi, K.S. 461
 Ravinder Reddy, P. 605
 Ray, Arindam 613
 Ray, Sudhabindu 371
 Rishi, Rahul 119
 Roy, Parthajit 623

 Sachan, Ajay Kumar 931
 Saha, Satadal 639, 649
 Saha, Soumyabrata 657
 Sai Hanuman, Akundi 669
 Saikumar, Tara 679
 Sailaja, M. 469
 Sameen Fatima, S. 413, 421

 Sandeep, Medepalli 715
 Sandhya, Nadella 703
 Santha Kumari, Seetala 309
 Saradhi, K. Sundeeep 923
 Saranya, A.A.S. 923
 Saravanakumar, G. 515
 Sarkar, Arup 731
 Sarkar, Ram 145, 739
 Sarkar, Subir Kumar 371
 Sasi Bhushana Rao, G. 363
 Satapathy, Suresh Chandra 29, 75, 265,
 337, 445, 453, 689, 819, 853, 923
 Satchidanandam, Anand 747
 Sathya Bama, S.B. 887
 Satyanarayana, P. 689
 Selvarani, R. 631
 Sengupta, Shampa 127
 Shaik Mohd., Zaheer Parvez 421
 Sharma, Shimantika 755
 Shashi, M. 897
 Shial, Rabindra Kumar 523
 Shirsat, Neeta 763
 Shiva, S. Chaitnya 405
 Si, Tapas 769
 Siddharth, 779
 Sil, Jaya 137, 281
 Singh, Arun Kumar 787
 Singh, Krishna 795
 Singh, Surendra 795
 Sinha Roy, Diptendu 523
 Sita Rama Murty, P. 469
 Siva Prasad, Kondapalli 345
 Siva Ranjani, Reddi 803
 Sivaselvan, K. 329
 Sonali, Sampangi 845
 Sowjanya, A.M. 337
 Srinivasa Murthy, Y.V. 345, 923
 Srinivasa Rao, Ch. 345
 Srinivasa Rao, D. 187
 Srinivasa Rao, Kandula 309
 Srinivasa Reddy, E. 869
 Srinivasu, Pakkurthi 819
 SriTeja Ayayangar V., R.K. 169
 Srivastava, Praveen Ranjan 827
 Srivastava, Priyanka 245
 Sriwastava, Brijesh Kumar 837
 Stephen, M. James 845, 853
 Subramani, C. 861
 Suma, V. 461
 Suman, Maloji 379

- Sumathi, R. 255
 Suresh, Ch. 853
 Suresh, G. Vijay 869
 Suvarna Kumar, Gogula 227
 Swamynathan, S. 539

 Tamil Selvi, S. 153
 Thakar, Urjita 497, 879
 Thamarai, S.M. 429
 Tibarewala, Dewaki N. 195
 Tudu, Bhimsen 387

 Uma Devi, T. 169

 Vaddi, Radhe Syam 353
 Vaidehi, M. 461

 Valarmathi, K. 515
 Vamsi Krishna, T.V.N.N.M. 187
 Vankayalapati, Hima Deepthi 353
 Vemulapalli, Saritha 897
 Venkateswara Reddy, E. 869
 Vernekar, Sumeet S. 907
 Vijayalakshmi Seshathri, C. 329, 723
 Vinay, A. 47
 Vinaya Babu, A. 669, 915
 Virajitha, Kota 353
 Voola, Persis 915

 Yadav, Anil Kumar 931
 Yenduri, Sumanth 179