Advances in Intelligent Systems and Computing 564

Khalid Saeed
Nabendu Chaki
Bibudhendu Pati
Sambit Bakshi
Durga Prasad Mohapatra *Editors*

Progress in Advanced Computing and Intelligent Engineering

Proceedings of ICACIE 2016, Volume 2

Springer

Advances in Intelligent Systems and Computing

Volume 564

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

The series "Advances in Intelligent Systems and Computing" contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing.

The publications within "Advances in Intelligent Systems and Computing" are primarily textbooks and proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

Khalid Saeed · Nabendu Chaki
Bibudhendu Pati · Sambit Bakshi
Durga Prasad Mohapatra
Editors

Progress in Advanced Computing and Intelligent Engineering

Proceedings of ICACIE 2016, Volume 2

Springer

Editors
Khalid Saeed
Faculty of Computer Science
Bialystok University of Technology
Białystok
Poland

Nabendu Chaki
Department of Computer Science
and Engineering
University of Calcutta
Kolkata
India

Bibudhendu Pati
C. V. Raman College of Engineering
Bhubaneswar, Odisha
India

Sambit Bakshi
Department of Computer Science
and Engineering
National Institute of Technology, Rourkela
Rourkela, Odisha
India

Durga Prasad Mohapatra
Department of Computer Science
and Engineering
National Institute of Technology, Rourkela
Rourkela, Odisha
India

ISSN 2194-5357 ISSN 2194-5365 (electronic)
Advances in Intelligent Systems and Computing
ISBN 978-981-10-6874-4 ISBN 978-981-10-6875-1 (eBook)
https://doi.org/10.1007/978-981-10-6875-1

Library of Congress Control Number: 2017955277

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

This volume contains the papers presented at International Conference on Advanced Computing and Intelligent Engineering (ICACIE 2016) that was held during December 23–25, 2016, at the C. V. Raman College of Engineering, Bhubaneswar, India (www.icacie.com). There were 638 submissions and each qualified submission was reviewed by a minimum of two Technical Program Committee members using the criteria of relevance, originality, technical quality, and presentation. The committee accepted and published in proceedings 136 full papers for oral presentation at the conference and the overall acceptance rate is 21.32%.

ICACIE is an initiative focusing on research and applications on several topics of advanced computing and intelligent engineering. The focus was also to present state-of-the-art scientific results, disseminate modern technologies, and promote collaborative research in advanced computing and intelligent engineering.

The accepted papers were chosen based on their research excellence, presentation quality, novelty, and the state-of-the-art representation. Researchers presented their work and had an excellent opportunity to interact with eminent professors and scholars in their area of research. All participants benefitted from discussions that facilitated the emergence of innovative ideas and approaches. Many distinguished professors, well-known scholars, industry leaders, and young researchers participated in making ICACIE 2016 an immense success.

We organized a special session named as Women in Engineering (WiE) on the topic "Empowerment of Women in the field of Engineering and Management" to encourage young women in the field of engineering and management to participate in the discussion. We had also industry and academia panel discussion and we invited people from software industries like TCS and Infosys.

We thank the Technical Program Committee members and all reviewers/sub-reviewers for their timely and thorough participation in the reviewing process.

We express our sincere gratitude to Shri Sanjib Kumar Rout, Chairman, C. V. Raman Group of Institutions, for allowing us to organize ICACIE 2016 on the campus. We also thank Prof. B. Bhattacharya, Principal, C. V. Raman College of Engineering, for his moral support. We thank Dr. Manmath Narayan Sahoo, NIT

Rourkela, Program Chair, for his valuable and timely support. We especially thank Dr. Chhabi Rani Panigrahi, C. V. Raman College of Engineering, for her support in local arrangement to make ICACIE 2016 a grand success. We appreciate the time and efforts put in by the members of the local organizing team at C. V. Raman College of Engineering, Bhubaneswar, especially the student volunteers, administrative staff, account section staff, and hostel management staff, who dedicated their time and efforts to ICACIE 2016. We thank Mr. Swagat Ranjan Sahoo for designing and maintaining ICACIE 2016 website.

We are very grateful to all our sponsors, especially DRDO and other local supporters, for their generous support toward ICACIE 2016.

Finally, we acknowledge the help of EasyChair in the submission, review, and proceedings creation processes. We are very pleased to express our sincere thanks to Springer, especially Mr. Anil Chandy, Mr. Harmen van Paradijs, Mr. Aninda Bose, and the editorial staff, for their support in publishing the proceedings of ICACIE 2016.

Białystok, Poland — Khalid Saeed
Kolkata, India — Nabendu Chaki
Bhubaneswar, India — Bibudhendu Pati
Rourkela, India — Sambit Bakshi
Rourkela, India — Durga Prasad Mohapatra

Organizing Committee

Advisory Board

Laxmi Narayan Bhuyan, FIEEE, FACM, FAAAS, University of California, Riverside, USA
Shyam Sundar Pattnaik, SMIEEE, FIETE, Biju Patnaik University of Technology, Odisha, India
Israel Koren, FIEEE, University of Massachusetts, USA
Katina Michael, SMIEEE, University of Wollongong, Australia
L.M. Patnaik, FIEEE, FINSA, FIETE, FIE, Indian Institute of Science, India
Rajib Mall, SMIEEE, Indian Institute of Technology Kharagpur, India
Prasant Mohapatra, FIEEE, University of California, USA
Abhay Bansal, SMIEEE, FIETE, FIET, Amity School of Engineering and Technology, India
Arun Somani, FIEEE, Iowa State University, USA
Atulya Nagar, Liverpool Hope University, UK
Brijesh Verma, SMIEEE, Central Queensland University, Australia
Debajyoti Mukhopadhyay, SMIEEE, FIE, Maharashtra Institute of Technology, India
George A. Tsihrintzis, University of Piraeus, Greece
Hugo Proenca, SMIEEE, University of Beira Interior, Portugal
Janusz Kacprzyk, FIEEE, Polish Academy of Sciences, Poland
Kenji Suzuki, SMIEEE, The University of Chicago, USA
Khalid Saeed, SMIEEE, AGH University of Science and Technology, Poland
Klaus David, University of Kassel, Germany
Gautam Das, FIEEE, University of Texas at Arlington, USA
Ganapati Panda, SMIEEE, IIT Bhubaneswar, India
Nabanita Das, SMIEEE, Indian Statistical Institute, Kolkata, India
Rama Krishna Challa, SMIEEE, NITTTR, Chandigarh, India
Biswanath Mukherjee, FIEEE, University of California, Davis, USA
Subhankar Dhar, FIEEE, San Jose State University, USA

Ashutosh Dutta, SMIEEE, AT&T Lab, USA
Kuan-Ching Li, FIET, SMIEEE, Providence University, Taiwan
Maode Ma, FIET, SMIEEE, Nanyang Technological University, Singapore
Massimo Tistarelli, FIAPR, SMIEEE, University of Sassari, Italy
Mohammad S. Obaidat, FIEEE, Monmouth University, USA
Sudip Misra, SMIEEE, Indian Institute of Technology Kharagpur, India
Michele Nappi, University of Salerno, Italy
Nishchal K. Verma, SMIEEE, Indian Institute of Technology Kanpur, India
Ouri E. Wolfson, FIEEE, FACM, University of Illinois at Chicago, USA
Pascal Lorenz, SMIEEE, FIARIA, University of Haute Alsace, France
Pierre Borne, FIEEE, Central School of Lille, France
Raj Jain, FIEEE, FACM, FAAAS, Washington University in St. Louis, USA
Rajkumar Buyya, SMIEEE, LMACM, The University of Melbourne, Australia
Raouf Boutaba, FIEEE, University of Waterloo, Canada
Saman Halgamuge, SMIEEE, University of Melbourne, Australia
Sansanee Auephanwiriyakul, SMIEEE, Chiang Mai University, Thailand
Subhash Saini, The National Aeronautics and Space Administration (NASA), USA
Arun Pujari, SMIEEE, Central University of Rajasthan, India
Sudhir Dixit, FIEEE, HP Lab, USA
Sanjay Mohapatra, Vice President, CSI, India

Chief Patron

Shri. Sanjib Kumar Rout, Chairman, C. V. Raman Group of Institutions, India

Patron

Smt. Shailja Rout, Managing Director, SSEPL Skills Pvt. Ltd, Odisha, India

Honorary General Chairs

Prasant Mohapatra, University of California, Davis, USA
Rajib Mall, Indian Institute of Technology Kharagpur, India
Sudip Misra, Indian Institute of Technology Kharagpur, India

Steering Committee

Kartik Chandra Patra, C. V. Raman College of Engineering, Odisha, India
Bhabes Bhattacharya, C. V. Raman College of Engineering, Odisha, India
Debdas Mishra, C. V. Raman College of Engineering, Odisha, India

General Chairs

Bibudhendu Pati, C. V. Raman College of Engineering, Odisha, India
Pankaj K. Sa, National Institute of Technology Rourkela, Odisha, India

Organizing Chairs

Chhabi Rani Panigrahi, C. V. Raman College of Engineering, Odisha, India
Sambit Bakshi, National Institute of Technology Rourkela, Odisha, India

Special Session Chairs

Rachita Mishra, C. V. Raman College of Engineering, Odisha, India
Brojo Kishore Mishra, C. V. Raman College of Engineering, Odisha, India

Program Chairs

Manmath Narayan Sahoo, National Institute of Technology Rourkela, Odisha, India
Subhas Chandra Misra, Indian Institute of Technology Kanpur, India

Publication Chairs

Sukant Kishoro Bisoy, C. V. Raman College of Engineering, Odisha, India
Soubhagya S. Barpanda, C. V. Raman College of Engineering, Odisha, India

Finance Chair

Mohit Ranjan Panda, C. V. Raman College of Engineering, Odisha, India

Website Chair

Swagat Ranjan Sahoo, C. V. Raman College of Engineering, Odisha, India

Registration Chair

Priyadarshini Nayak, C. V. Raman College of Engineering, Odisha, India

Publicity Chair

Tanmay Kumar Das, C. V. Raman College of Engineering, Odisha, India

Organizing Committee

Amardeep Das
Abhaya Kumar Sahoo
Amrut Ranjan Jena
Amulya Kumar Satpathy
Babitarani Garanayak
Banee Bandana Das
Bijaylaxmi Panda
Biswajit Upadhyay
Chhabirani Mohapatra
Chandra kanta Mohanty
Debasis Mohanty
Debapriya Panda
Harapriya Rout
Himansu Das
Jyotiranjan Swain
Kartik chandra Jena
Khitish Kumar Gadnayak
Lalat Kishore Choudhury
M. Priyattama Sahoo
Madhusmita Mishra
Mamata Rani Das
Mamata Rath
Manas Ranjan Mishra
Monalisa Mishra
Nilamadhaba Dash
Prakash Chandra Sahu
Prashanta Kumar Dash
Rashmiprava Sahoo
Rojalin Priyadarshini
Sharmistha Puhan
Sasmita Parida
Satyashree Samal
Soumya Sahoo
Shreela Dash
Sujit Mohapatra
Sunil Kumar Mohapatra

Sushruta Mishra
Suvendu Chandan Nayak

Technical Programme Committee

Chui Kwok Tai, City University of Hong Kong, Hong Kong
Bernd E. Wolfinger, University of Hamburg, Hamburg
Amin Al-Habaibeh, Nottingham Trent University, UK
Carlo Vallati, University of Pisa, Italy
Rajendra Prasath, University College Cork, Ireland
Chi-Wai Chow, National Chiao Tung University, Taiwan
Mohammed Ghazal, Abu Dhabi University, UAE
Felix Albu, Valahia University of Targoviste, Romania
Vasanth Iyer, Florida International University, USA
Victor Govindaswaormy, Concordia University Chicago, USA
Priyadarshi Kanungo, C. V. Raman College of Engineering, Odisha, India
Sangram Mohapatra, C. V. Raman College of Engineering, Odisha, India
Saikat Charjee, C. V. Raman College of Engineering, Odisha, India
Chakchai So-In, Khon Kaen University, Thailand
Cristina Alcaraz, University of Malaga, Spain
Barun Kumar Saha, Indian Institute of Technology Kharagpur, India
Pushpendu Kar, Nanyang Technological University, Singapore
Samaresh Bera, Indian Institute of Technology Kharagpur, India
Ayan Mandal, Indian Institute of Technology Kharagpur, India
Tamoghna Ojha, Indian Institute of Technology Kharagpur, India
Subhadeep Sarkar, Indian Institute of Technology Kharagpur, India
Somanath Tripathy, Indian Institute of Technology Patna, India
George Caridakis, University of the Aegean, Greece
Carlos Alberto Malcher Bastos, Universidade Federal Fluminense, Brazil
Laizhong Cui, Shenzhen University, China
Srinivas Prasad, GMRIT, Rajam, India
Prasant Kumar Sahu, Indian Institute of Technology Bhubaneswar, India
Mohand Lagha, University of Blida, Algeria
Vincenzo Eramo, University of Rome, La Sapienza, Italy
Ruggero Donida Labati, Università degli Studi di Milano, Italy
Satyananda Rai, SIT, Bhubaneswar, India
Dinesh Bhatia, North Eastern Hill University, Meghalaya, India
Vasilis Friderikos, King's College London, UK
C. Lakshmi Devasena, IFHE University, India
Arijit Roy, Indian Institute of Technology Kharagpur, India
Roberto Caldelli, Università degli Studi Firenze, Italy
Christos Bouras, University of Patras, Greece

Iti Saha Misra, Jadavpur University, India
Salil Kumar Sanyal, Jadavpur University, India
J. Joshua Thomas, School of Engineering, KDU Penang University College, Penang
Shibendu Debbarma, Tripura University, India
Angelo Genovese, Università degli Studi di Milano, Italy
Marco Mussetta, Politecnico Di Milano, Italy
Radu-Emil Precup, Politehnica University of Timisoara, Romania
Debi Acharjya, VIT University, Vellore, India
Samaresh Mishra, KIIT University, Bhubaneswar, India
Rio D'Souza, St Joseph Engineering College, Mangalore, India
Yogesh Dandawate, Vishwakarma Institute of Information Technology, Pune, India
Sanjay Singh, Manipal Institute of Technology, Manipal, India
Rajesh R., Central University of Kerala, India
Abhishek Ray, KIIT University, Bhubaneswar, India
Lalat Indu Giri, National Institute of Technology, Goa, India
Debdas Mishra, C. V. Raman College of Engineering, Odisha, India
Ameresh Panda, C. V. Raman College of Engineering, Odisha, India
Tripti Swarnakar, SOA University, Bhubaneswar, India
Judhistir Mohapatro, National Institute of Technology, Delhi, India
Manas Khatua, SUTD, Singapore
Sujata Pal, University of Waterloo, Canada
Sumit Goswami, DRDO, New Delhi, India
Rabi Narayana Sathpathy, HIT, Bhubaneswar, India
Harihar Kalia, SEC, India
Hari Saran Dash, Infosys, Bhubaneswar, India
Siba Kumar Udgata, University of Hyderabad, India
Mu-Song Chen, Da-Yeh University, Taiwan
Félix J. García, University of Murcia, Spain
Prasant Kumar Pattnaik, KIIT University, India
Poornalatha G., MIT, Manipal, India
Nishant Doshi, MEFGI, Rajkot, India
V.N. Manjunath Aradhya, JCE, Mysore, India
Prabhakar C.J., Kuvempu University, Karnataka, India
Enrico Cambiaso, National Research Council, CNR-IEIIT, Italy
Gianluigi Ferrari, University of Parma, Italy
Elena Benderskaya, Saint-Petersburg State Politechnical University, Russia
Josep Domènech, Universitat Politècnica de València, Spain
Himansu Das, KIIT University, India
Vivek Kumar Sehagl, Jaypee University of Information Technology, Waknaghat, India
Monish Chatterjee, Asansol Engineering College, Asansol, India
Teresa Gomes, Universidade de Coimbra—Polo II, Portugal
Chandralekha, DRIEMS, India
Haoxiang Wang, Cornell University, USA

Contents

About the Editors

Khalid Saeed received B.Sc. degree in Electrical and Electronics Engineering in 1976 from Baghdad University, M.Sc. and Ph.D. degrees from Wroclaw University of Technology, in Poland in 1978 and 1981, respectively. He received his D.Sc. degree (Habilitation) in Computer Science from Polish Academy of Sciences in Warsaw in 2007. He is Professor of Computer Science at AGH University of Science and Technology in Poland. He has authored more than 200 publications including 27 edited books, journals, and conference proceedings, and 8 text and reference books. He supervised more than 110 M.Sc. and 12 Ph.D. theses. His areas of interest are Biometrics, Image Analysis, and Processing and Computer Information Systems. He gave 45 invited lectures and keynotes in different universities in Europe, China, India, South Korea, and Japan. The talks were on Biometric Image Processing and Analysis. He received about 19 academic awards. Khalid Saeed is a member of the editorial boards of over 15 international journals and conferences. He is IEEE Senior Member and has been selected as IEEE Distinguished Speaker for 2011–2013 and 2014–2016. Khalid Saeed is Editor-in-Chief of International Journal of Biometrics with Inderscience Publishers.

Nabendu Chaki is Professor in the Department Computer Science and Engineering, University of Calcutta, Kolkata, India. Dr. Chaki did his first graduation in Physics from the legendary Presidency College in Kolkata and then in Computer Science and Engineering, University of Calcutta. He completed Ph.D. in 2000 from Jadavpur University, India. He is sharing two US patents and one patent in Japan with his students. Professor Chaki is quite active in developing international standards for Software Engineering. He represents the country in the Global Directory (GD) for ISO-IEC. Besides editing more than 20 books in different Springer series including LNCS, Dr. Chaki has authored 5 text and research books and about 130 peer-reviewed research papers in journals and international conferences. His areas of research interests include distributed computing, image processing, and software engineering. Dr. Chaki has served as Research Assistant Professor in the Ph.D. program in Software Engineering in U.S. Naval Postgraduate School, Monterey, CA. He is having strong and active collaborations in US,

Europe, Australia, and other institutes and industries in India. He is visiting faculty member for many universities in India and abroad. Dr. Chaki has been the Knowledge Area Editor in Mathematical Foundation for the SWEBOK project of the IEEE Computer Society. Besides being in the editorial board for several international journals, he has also served in the committees of more than 50 international conferences. Professor Chaki is the founder Chair of ACM Professional Chapter in Kolkata.

Bibudhendu Pati is Associate Professor in the Department of Computer Science and Engineering at C. V. Raman College of Engineering, Bhubaneswar, Odisha, India. Dr. Pati has total 19 Years of Experience in Teaching, Research, and Industry. His interest areas include Wireless Sensor Networks, Cloud Computing, Big Data, Internet of Things, and Network Virtualization. He completed his Ph.D. from IIT Kharagpur in 2014, MBA from Punjab Technological University in 2010, and M.E. from NITTTR, Chandigarh in 2008. He is Life Member of Indian Society of Technical Education (ISTE), Member of IEEE, ACM, CSI and Computer Science and Engineering Research Group, IIT Kharagpur. He has got several papers published in journals, conference proceedings, and books.

Sambit Bakshi is Assistant Professor in the Department of Computer Science and Engineering at National Institute of Technology Rourkela, Odisha, India. He completed his M.Tech. and Ph.D. from NIT Rourkela in 2011 and 2014, respectively. His research interest areas are Biometric Security and Visual Surveillance. He has several journal publications, book chapters, two authored books, and six edited volumes to his credit. He has been teaching subjects like biometric security, statistical analysis, linear algebra and statistical analysis laboratory, digital image processing, etc. He has also been involved in many professional and editorial activities.

Durga Prasad Mohapatra received his Ph.D. from Indian Institute of Technology Kharagpur, India. He joined the Department of Computer Science and Engineering at the National Institute of Technology, Rourkela, India in 1996, where he is presently serving as Associate Professor. His research interests include Software Engineering, Real-Time Systems, Discrete Mathematics, and Distributed Computing. He has published over 30 research papers in these fields in various international journals and conferences. He has received several project grants from DST and UGC, Government of India. He has received the Young Scientist Award for the year 2006 by Orissa Bigyan Academy. He has also received the Prof. K. Arumugam National Award and the Maharashtra State National Award for outstanding research work in Software Engineering for the years 2009 and 2010, respectively, from the Indian Society for Technical Education (ISTE), New Delhi. He is going to receive the Bharat Sikshya Ratan Award for significant contribution in academics awarded by the Global Society for Health and Educational Growth, Delhi.

Part I
Learning Algorithms, Neural Networks and Pattern Recognition

A Framework to Enhance the Learning Outcome with Fuzzy Logic-Based ABLS (Adaptive Behaviourial Learning System)

Suman Deb, Jagrati and Paritosh Bhattacharya

Abstract Intelligent adaptive learning style in education system is a demanding trend in expertise learning with a specific outcome. Taking advantage of the continuous improving learning system for teaching purpose increases the students learning ability. Learning style recommends the mode in which one understands and wants to learn. The proposed method clusters the students of a class according to individual's natural learning ability. It gives the clear association and definition to each member belonging to a particular cluster. It is an enhanced design of deliverable for providing enhanced and effective outcome which the teacher can customize for the class as well as for their teaching methodologies. Experiments show that the proposed system can significantly help the teacher in predetermining the expectation, level of understanding, expandability, etc. With periodic outcome from the class evaluation, the teacher can steer the teaching learning process. This can be quantified dynamically and motivates the learners in a continuous process of teaching learning method.

Keywords Learning style · Learner's profile · Fuzzy inference engine
Clustering ABLS

S. Deb · Jagrati · P. Bhattacharya (✉)
Computer Science and Engineering Department,
National Institute of Technology, Agartala, India
e-mail: pari76@rediffmail.com

S. Deb
e-mail: sumandebcs@springer.com

Jagrati
e-mail: jagratimahalwal1992@springer.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_1

1 Introduction

Education could not remain passive and nonchalant; the traditional teaching learning process should be revised and reconsidered to upsurge the learning outcome. The advent of high-quality teaching is a result of the incompetence of traditional courses to capture the specifics of the course achieved by the learner. Teaching learning process is an interactive process where teacher donates the knowledge and student accepts it, and to enhance this pedagogy the acceptor's preference in gaining knowledge should be considered. The practice of learning in preferred natural learning manner results in high performance and thus helps in achieving the goal of enhanced learning outcomes. Learning styles are intellectual, emotional and behavioural characteristics that help in indicating how a leaner concludes, discovers and acknowledges the pedagogy. Depending upon learner's interest, we can identify the learning style and enhance their learning by gradually adjusting the teaching methodologies. The analysis and the interpretation of learners behaviour is very vital factor in personalizing the teaching learning process among the class of students having heterogeneous learning cultures. In our approach, we are determining the learner's interest through a pre-evaluation process, clustering them according to those interest. The pre-assessment performed as per their behavioural learning styles using Honey and Mumford questionnaire, depicting particulars learning techniques. Honey and Mumford questionnaire [1] outcome results in four clusters namely Activists, Reflectors, Theorists and Pragmatists, at the beginning of the course. These groups will help in observing the diversity among the same learning culture. On further analysis by applying fuzzy inference system, we denote the degree of membership of individual in the cluster.

2 Survey on Learning Styles

Over the years, various learning style models have been developed by theorists and researchers. They have contributed valuable resolutions in categorizing them into groups according to their theories. David Kolb [2] published his learning style model in 1984 from which he developed his learning style inventory. Kolb's model outlines two related approaches towards grasping experience: Concrete Experience and Abstract Conceptualization, as well as two related approaches towards transforming experience: Reflective Observation and Active Experimentation. The distinct models presented are based on different learning theories [3] such as Kolb's model is an experimental theory model in which learners are categorized as Divergers, Assimilators and Convergers; Honey and Mumford model is a behavioural theory model and learners are categorized as Activists, Theorists, Pragmatists and Reflectors; Gregorc model [4] is cognitive theory model, where learners are classified as Abstract sequential, Abstract random, Concrete sequential and Concrete random; Carl and Myers Briggs Indicator model is based on person's personality, where learners

are classified as Extroversion, Introversion, Sensing, Intuitions, Thinking, Feeling, Judgement and Perception; Felder-Silverman [5] model is a physiological theory model, where learners are classified as Sensing–Intuitive, Visual–Verbal, Indicative–Deductive, Active, Reflective and Sequential global.

The focus of this study is on academic learning in which the clustering is based on the Honey and Mumford [1] learning questionnaire model. Peter Honey and Alan Mumford adapted Kolb's experimental learning model for their learning cycle theory. Honey and Mumford presented a self-development tool consisting of 80 numbers of questionnaires which are to be answered YES or NO and get one point for each YES. The evaluation results in four groups namely Activists, Reflectors, Theorists and Pragmatists.

- Activists are those students who learn by participating in new experiments and activities. They are enthusiastic about learning and implementing them.
- Reflectors understand the problem thoroughly and create their viewpoint before engaging themselves or coming to any conclusion.
- Theorists involve themselves in observations and try to integrate them with logical theories. They perform the step-by-step evolution of any given situation.
- Pragmatists apply their conceptual theories to the problems and generate profound results. They are keen to try things.

3 Motivation

Many attempts have been made [6] to enhance students learning outcome and hence academic achievements. The main motive of teaching learning process is to develop the student's mind so that the student can stand and gain success in the society. One of the critical factors that affect the teaching and learning process is a learning style. This is to design an effective instruction that can help in acknowledging the diverse learning style among the students. This type of nontraditional learning paradigms requires a fundamental shift in the classroom input to achieve benchmarks. This shifts from pedagogy that is centred on providing instruction to the one that focuses on learning. Our system helps the students as well as teachers to define the heterogeneous objectives among homogeneous groups, thereby increasing the learning outcome.

4 System Architecture

4.1 General

The general architecture of the proposed system consists of two main components that form stand-alone objects while simultaneously proving to be interdependent. The first object is a learner or a group of learners participating individually as well

in collaborative learning and assessing process. The root of the system is formed on this entity, and the working process of the second entity functionally depends on the first. The second entity is the guide or experts whose goal is to design and operate the whole system for the better computational and educational efficiency. This group defines the content of course and forms the objective as per learner's knowledge level in the specific learning subject. This group of experts maintains the whole system in three steps, that is, capturing the data, analysing the data and measuring the outcome of the individuals and hence enhancing efficiency of learning outcome.

4.2 Capture

This module will enable the teacher to cluster the students of a class on the basis of a pre-assessment using Honey and Mumford learning style questionnaire model. Students were provided with a number of questionnaires in which they have to select YES or NO. Individual results were calculated; one point for the question for which the student has marked YES and no point for the answered marked as NO or even left unattempt, and according to their score they were assigned to groups to which students are more similar to each other than to those who are in other groups. Then according to their grasping potential, the group is assigned with their respective objectives. The group was then assessed and the results are stored in database for the further analysis (Fig. 1).

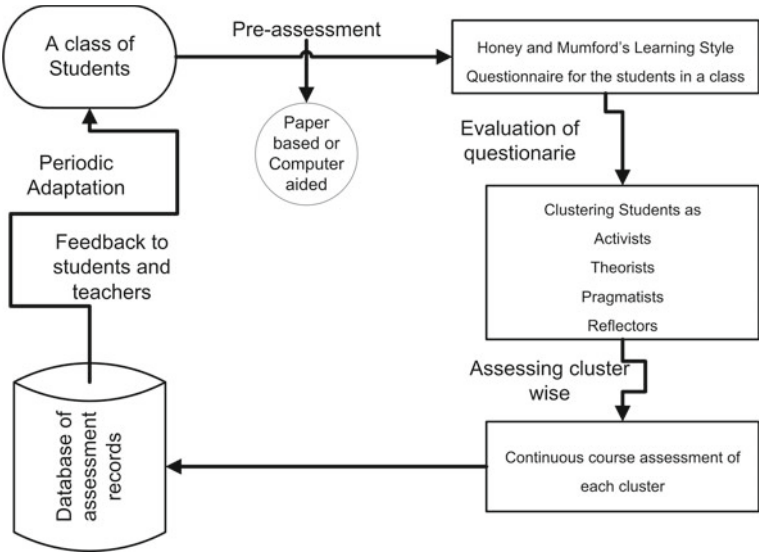


Fig. 1 Modules of student clustering using Honey and Mumford learning style questionnaire model

4.3 Dataset

The dataset used in this system is a two-dimensional vector consisting of accuracy and the time taken by the each student. The accuracy is calculated as the total marks obtained to the maximum marks obtained. This dataset is analysed using fuzzy logic controller which provides the membership for each individual among the cluster.

4.4 Analysis

This module will help in assessing each cluster as well as the student of the particular cluster individually.

STEP1: Fuzzification [7] of these input values to fuzzy values is done by using Gaussian membership function:

$$f(x; \sigma_i, c) = exp(-(x - c)^2/2\sigma_i^2), \tag{1}$$

where c is the centre (i.e. mean), and σ_i is the width, i.e. standard deviation of ith fuzzy set. The Gaussian membership function is more robust than the triangular as it uses two parameters instead of three, thereby reducing the degree of freedom. In our system, we choose the three centres, i.e. 0.2, 0.5 and 0.8. Therefore, Gaussian

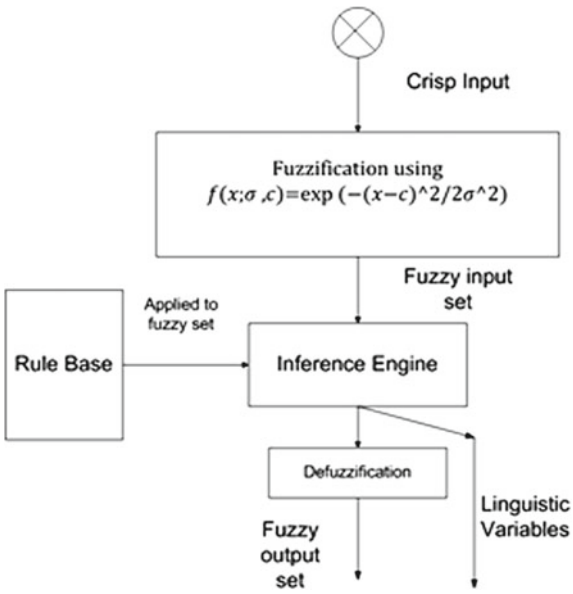


Fig. 2 Fuzzy logic controller

membership functions increase the probability that every rule fires from the inference engine.

STEP 2: Inference-based system [6] with the help of defined IF-THEN rules in the rule base. The rules in rule base are based on the accuracy matrix and the time rate matrix of individual student and the linguistic variables are assigned to each, respectively (Fig. 2).

STEP 3: Defuzzification is a process which results in crisp output from given fuzzy sets. In our work, we have used the Centre of Gravity method or COG. It gives the crisp value for the centroid area of the curve. The output membership functions to which the fuzzy outputs are transposed are restricted to being singletons:

$$Crispoutput = (fuzzyoutput) \times (singletonvalueonx - axis)/(fuzzyoutput). \quad (2)$$

4.5 Outcome

The Mamdani's fuzzy inference method produces the fuzzy output set and relative linguistic variables to define the students of a cluster. For example, in a cluster, if student A obtains 9 marks out of 10, then the time taken was more than the assigned time. Another student B is getting 9 marks out of 10 and the time taken is less than student A, then the degree of membership of each student belonging to the same cluster will be different. The output from the second module, i.e. Analysis, provides an in-depth of individual's performance belonging to different clusters (Fig. 3).

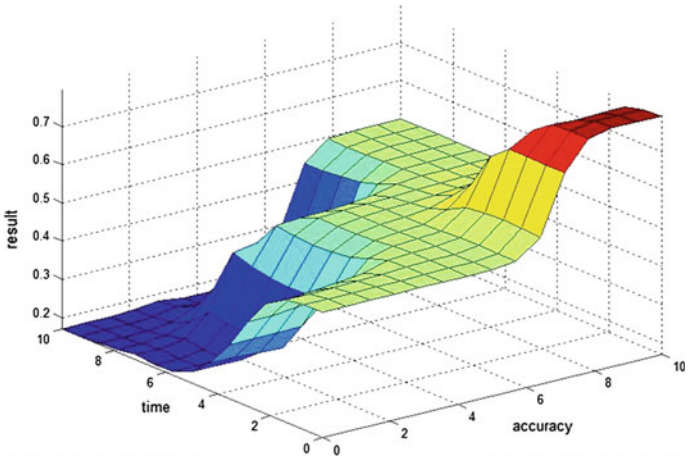


Fig. 3 Fuzzy surface view of rule base

A performance graph for each group plotted is studied. The graph (Fig. 4) helps the students in knowing their performance and membership in the cluster and motivates them to enhance it. The guide can also easily identify the grasping efficiency of a student belonging to a group and revise the teaching technique, if necessary, accordingly. This continuous process of clustering, assessing and revising is helpful

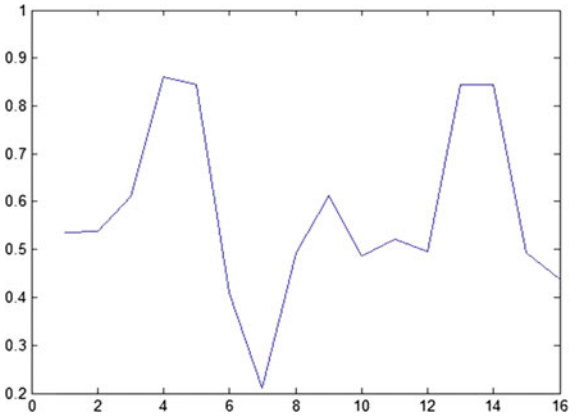


Fig. 4 Membership graph of students in one cluster

Table 1 Input dataset and result of students belonging to activist cluster

Sl no	Input		
	Accuracy	Time taken (min)	Result
1	7	10	0.5366
2	8	10	0.5387
3	6	9	0.6100
4	9	8	0.8603
5	8	9	0.8453
6	3	12	0.4086
7	2	12	0.2108
8	2	8	0.4946
9	6	7	0.6119
10	4	10	0.4873
11	10	10	0.5223
12	10	15	0.4948
13	10	7	0.8436
14	8	9	0.8453
15	9	12	0.4946
16	6	12	0.4383
17	7	9	0.7645
18	6	10	0.5268

in increasing the learning outcome of the overall class. It will also help in enhancing the learning–retaining–revising process (Table 1).

5 Adaptation

The system helps both the teacher and student in unique ways. The overall system is able to continuously adapt according to its learners' preferences, history and specific needs as they emerge during its usage. The system stores the learners profile in database and upgrades any changes that occur to it during the course. The profile includes the preferred learning style, answers of the questionnaire(s) conducted during the entire course and the assessment of each. The different modules of the system provide detailed information of student's performance throughout the year. The students profile obtained from the system helps the guide to filter or collocate the students in appropriate groups and also helps in providing an accurate feedback.

6 Conclusion

The proposed system introduced as novel integrated learning process. It is able to continuously adapt according to its learners' preferences, history and specific needs as they emerge during its usage in a course. The different modules of the system provide detailed information of student's performance throughout the year. The system stores the learners profile in database and upgrades any changes that occur to it during the course. This profile allows the students in upgrading their learning and enhancing the learning outcome by assessing the feedback provided by the system. The student's profile also helps the guide to filter or collocate the students in the appropriate groups and gradually design their teaching methodologies so as to acknowledge diverse learning among student.

Declaration Authors have obtained all ethical approvals from appropriate ethical committee and approval from the students or from their parents/LAR (because the students are minor) who participated in this study.

References

1. Honey, P., Mumford, A.: The Learning Styles Helper's Guide. Peter Honey Learning (2000)
2. Experiential Learning. Google Books (2017). https://books.google.co.in/books?l=en&lr=&id=jpbeBQAAQBAJ&oi=fnd&pg=. Accessed 12 July 2017
3. Deborah, L.J., et al.: Fuzzy-Logic Based Learning Style Prediction in e-learning Using Web Interface Information, vol. 40, Part 2, pp. 379–394. Indian Academy of Sciences (2015)

4. Sewall, T.: The Measurement of Learning Style: A Critique of Four Assessment Tools (1986). https://eric.ed.gov/?id=ED267247. Accessed 12 July 2017
5. Viola, S.R., Timmaso, L.: In-depth analysis of Felder-Silverman learning style dimensions. J. Res. Technol. Educ. **40**(1)
6. Agbonifo, O.C.: Fuzzy c-means clustering model for identification of students learning preferences in online environment. IJCAIT **4** (2013)
7. Fuzzy Modelling and Control by Andrzej Piegat
8. Fuzzy Logic Toolbox 2.2.7. http://www.mathworks.com/products/fuzzylogic

Experimental Comparison of Sampling Techniques for Imbalanced Datasets Using Various Classification Models

Sanjibani Sudha Pattanayak and Minakhi Rout

Abstract Imbalanced dataset is a dataset, in which the number of samples in different classes is highly uneven, which makes it very challenging for classification, i.e., classification becomes very tough as the result may get biased by the dominating class values. But misclassification of minor class sample or interested samples is very much costlier. So to provide solution to this problem, various studies have been made out of which sampling techniques are successfully adopted to preprocess the imbalance datasets. In this paper, experimental comparison of two pioneering sampling techniques SMOTE and MWMOTE is simulated using the classification models SVM, RBF, and MLP.

Keywords Sampling techniques · SMOTE · MWMOTE · SVM · RBF · MLP

1 Introduction

The dataset, in which the ratio of number of samples in major and minor classes is very high, is called imbalanced dataset. This imbalanced nature of dataset is undesirable for a good classification because classifier shows very good result for major datasets but poor result for the minor datasets, i.e., the classifier could not be trained for minor class properly. But misclassification of minor class sample is much more costlier than major class sample, i.e., classification of minor class sample with high accuracy is very much essential. Unfortunately, most of the real-world applications like fraudulent transaction detections, detecting network intrusion, Web mining, medical diagnostics, and many other find imbalanced data. And it is always very much essential to give justice to the minor class. Various solutions have been proposed till date to eradicate the imbalanced nature. The solutions might be in data

S. S. Pattanayak (✉) · M. Rout
ITER, Siksha O Anusandhan University, Bhubaneswar 751030, Odisha, India
e-mail: sanjibani@gmail.com

M. Rout
e-mail: minakhirout@soauniversity.ac.in

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_2

level or in algorithm level. At data level, sampling is considered as a major technique to handle imbalance nature of dataset. Data are sampled to bring balance between uneven classes. Broadly, the sampling techniques can be classified into two categories: undersampling and oversampling. Using undersampling, the size of majority class can be reduced to match with the minority class. But in this approach, there is a chance of losing important samples. So, in many cases, oversampling is adopted instead of undersampling, whereby generating new synthetic samples, the size of minority class can be developed.

1.1 Literature Survey

Significant works have been done to handle the imbalance nature of datasets. Sampling is one of the major techniques which have been adopted highly by researchers. In SMOTE (Chawla 2008), the dataset is oversampled by creating synthetic examples of minority class. Synthetic samples are generated by considering the feature sample (Minor sample) and its k-nearest neighbor. K value is chosen depending upon how many synthetic samples you need to generate. The difference between the featured sample and nearest neighbor is taken and it is multiplied with a random value in the range (0, 1). This result is added with the same featured sample to generate the synthetic sample. This is how it makes the region of minor samples more general [1]. Many variations of SMOTE have been mentioned here. Chawla et al. [2] proposed SMOTEboost that combines the features of SMOTE and boosting to focus more on difficult examples that belong to the minority class than to the majority class. It gives higher weight to synthetic samples and thus adjusts the imbalance nature [2]. Unlike SMOTE, which generates the synthetic samples from every minor sample, MSMOTE method by Hu et al. [3] considers the distribution of minority class samples and latent noise in the dataset. It eliminates those noisy samples [3]. Maciejewski and Stefanowski [4] proposed LNSMOTE which is more careful about choosing the local neighborhood to avoid the overgeneralization cases of SMOTE [4]. Han et al. [5] proposed the method of Borderline SMOTE, in which, instead of considering all the minor samples, only the examples on the borderline and the ones nearby are used to generate synthetic samples as they prone to more error, i.e., misclassification [5]. He et al. [6] proposed Adasyn algorithm that assigns weight to minor samples. As per the algorithm, the samples which are more difficult to learn will be assigned a higher weight and more synthetic samples will be generated from the samples having higher weight than others and this is how it tries to bring justification [6]. MWMOTE [7] first identifies the difficult to learn minor samples and assigns weight according to their distance from nearest majority class samples. It also eliminates the noisy samples. Then, it generates the synthetic samples by forming clusters of minor samples. Now, the featured sample and the nearest neighbor sample will be chosen from the same cluster. Hence, create more concrete synthetic samples [7]. Jayashree and Gavya [8] have used oversampling technique MWMOTE and undersampling technique SSO for better imbalanced learning [19].

1.2 Objective

The objective of this paper is to study the performance of classifiers on imbalanced dataset before oversampling and after oversampling. Also, the two well-known oversampling techniques, Synthetic Minority Oversampling Technique (SMOTE) and Majority Weighted Minority Oversampling Technique (MWMOTE), have been compared using three distinct classification models.

2 Classification Models and Sampling Techniques

2.1 Classification Models

In this paper, popular classification models like SVM and artificial neural networks like MLP and RBF have been used for simulation.

SVM: Support vector machine is a supervised learning machine used to classify two-class problems linearly. Multi-class problems can be modified or shaped to two-class problems so that SVM can work with that too. SVM operates by constructing a hyperplane between the two sets of data. It is very important to choose the right hyperplane for better accuracy because the data near the borderline are most difficult to learn. The hyperplane that has the largest distance to the nearest data point is considered as the best hyperplane. Generally, the kernel trick is used to solve the nonlinear problems. Different variations of the SVM have been developed and tested by the researchers for imbalanced datasets. Tang et al. [14] proposed modified SVM to handle imbalanced dataset and they found that among different variations of SVM, GSVM-RU shows the best efficiency [14]. Another variation of SVM, z-SVM of Imam et al. [15], maintains a good boundary between classes and the separator line [15].

RBFN: Radial basis function network is another powerful supervised learning machine which uses a radial basis function to train the neurons. It has a three-layer architecture in which the input layer consists of the feature vectors which will be classified. Hidden layer uses a radial basis function, whereas the output layer deals with weight vector to generate the actual output for the given input. Commonly, Gaussian function is used as the radial basis function. Perez-Godoy et al. [16] experimented on RBF with LMS and SVD for imbalanced dataset and found that SVD outperforms LMS [16]. Haddad et al. [17] studied the effect of imbalanced training set size using RBF and they found that with the increased size of imbalanced dataset, the classifier performance degrades [17].

MLP: Multilayer perceptron uses backpropagation technique to train the network in a supervised method. It has input layer, output layer, and in between number of hidden layers. It uses nonlinear activation functions to solve nonlinear problems. To get

a right match of the input and output, the weight vector that has been provided at the input layer is adjusted with the backpropagation of error to the neural network. Bruzzone and Serpico 1997 proposed a multilayer perceptron to classify the imbalanced remote-sensing data and found improved speed and more stable result [8]. Oh 2011 modified the error function of MLP to update the weight value. Here, the weight value of minority class intensifies and of majority class weakens, and hence avoids the biasness in the imbalanced datasets [18].

2.2 Sampling Techniques

Sampling techniques are most popular and widely used techniques to handle imbalanced dataset problem. Here, it tries to balance the number of samples by dropping few samples from majority class or adding new synthetic samples to minority class. Two popular oversampling methods SMOTE and MWMOTE have been discussed below. Instead of replication of the same data, if new synthetic samples will be generated, the region of minority class will become more broader, and hence learning can become more generic.

SMOTE: SMOTE which is abbreviated for Synthetic Minority Oversampling Technique generates synthetic samples by considering k-nearest neighbors of each of the minority samples. The value of k depends on how many number of synthetic samples will be generated. The difference between the feature sample under consideration and the neighboring sample will be computed which will be multiplied with a random number between 0 and 1. This multiplication result will be now added with feature sample in turn to create the synthetic sample.

MWMOTE: MWMOTE or Majority Weighted Minority Oversampling Technique is another oversampling technique which is more cautious in choosing the minority class samples to generate synthetic samples. Unlike SMOTE, it gives priority to those minority class samples which are more difficult to learn. Experiment shows that the samples which lie near border, the samples which are a part of a minority cluster that is sparsely populated, are more difficult to classify. So at first, all difficult to learn minority samples are identified. Next, their difficulty level is found out so that it can be decided which minority sample will contribute how many numbers of synthetic samples.

3 Performance Measure

Confusion Matrix Confusion matrix is a tabular representation of accuracy of a classifier. Rows represent the true class, whereas the columns represent the predicted class. TP stands for true positive, i.e., TP value says how many positive samples (i.e., minority class samples) are identified correctly. FP (False positive) indicates the number of positive samples misclassified. Similarly, TN value says the negative

samples (majority class samples) that are classified correctly and FN represents number of negative samples that are misclassified. So the main diagonal of the matrix shows how many samples are predicted correctly.

The traditional method to measure the accuracy, shown in Eq. 1, is not suitable for imbalanced dataset:

$$OverallAccuracy = \frac{(TP + TN)}{(TP + FN + FP + TN)}. \tag{1}$$

In imbalanced dataset, more than 98% data belong to majority class set. So even if the classifier failed to identify the minority class samples, it will show very high accuracy. But high accuracy with correct identification of minority class set is very much important. (For example, correct identification of one positive cancer sample is highly necessary among 1 lakh negative samples. The misclassification of the same is intolerably dangerous here.) Hence, various metrics listed from Eqs. 2–7 are used for performance measure of classifier over imbalanced datasets.

TPR says out of total minority class samples, how many are classified correctly:

$$TPR(TruePositiveRateorRecallorSensitivity) = \frac{TP}{(TP + FN)}. \tag{2}$$

TNR measures how many majority class samples are correctly classified. High TNR value is desirable for more majority class samples that are correctly identified:

$$TNR(TrueNegativeRateorSpecificity) = \frac{TN}{(TN + FP)}. \tag{3}$$

FPR or False Positive Rate indicates how many minority class samples are misclassified:

$$FPR = \frac{FP}{(TN + FP)}. \tag{4}$$

While TPR or recall says out of total actual minority class samples how many are classified correctly, precision says out of total predicted minority class samples, how many are actually belong to minority class:

$$Precision = \frac{TP}{(TP + FP)}. \tag{5}$$

Fmeasure is the harmonic mean of precision and recall, i.e., it gives the balance between precision and recall. Its value will be high for high precision and high recall:

$$Fmeasure = \frac{(2 * Precision * Recall)}{(Precision + Recall)}. \tag{6}$$

Table 1 Dataset description

Dataset name	No. of features	Total no. of samples	No. of samples in majority class	No. of samples in minority class	Imbalance ratio
Wineequality	11	691	681	10	68.1
Poker	10	1477	1460	17	85.88
kddcup-rootkit-imap-vs	41	2225	2203	22	100.14

Gmean is the geometric mean, i.e., a kind of average of two parameters recall or sensitivity and specificity:

$$Gmean = \sqrt{Sensitivity * Specificity}. \tag{7}$$

4 Dataset Description

In this paper, the simulation has been done on the openly available two-class datasets.

The datasets' descriptions have been given in Table 1. Wineequality dataset has total 691 samples out of which 681 samples belong to majority class, whereas only 10 samples belong to minority class. Hence, the imbalance ratio is here 68.1. It poses 11 numbers of feature. Next dataset, i.e., Poker which is having 10 features, is a bigger dataset than the first one. It has 1460 samples in majority class and only 17 samples in minority class. So here the imbalance ratio is 85.88. The third dataset kddcup is the biggest one among three having more features and more imbalanced proportion of majority and minority class samples. Here, out of total 2225 samples, 2203 samples belong to majority class and only 22 samples belong to minority class. Hence, the imbalance ratio is more than 100.

5 Simulation Study and Result

Different steps that are involved in the entire simulation process are shown below:

- Choosing the Datasets,
- Introducing Synthetic Samples,
- Modeling the Classifiers,
- Preparing the Training data and Testing data, and
- Measuring the performance.

Without oversampling, when the datasets are passed to these classifiers, it shows very poor performance, i.e., the classifiers could not be trained for the minority class. So for correct classification, datasets are oversampled to generate synthetic samples

of minority class. For each dataset, two sets of oversampled data are generated using SMOTE and using MWMOTE. As per previous studies, 200–400% sampling shows good result, so here the dataset has been sampled to 400%.

Then the oversampled data are passed to the classifiers. Each dataset has been distributed in 80:20 ratios for training and testing purpose. The outcome of each classification model has been shown in separate table. Here, the simulation has been done using the mathematical tool Matlab. The underperformance of classifiers without sampling is very clear in Table 2. After sampling, significant improvement in the performance is observed which has been shown in Tables 3, 4 and 5.

Among the three classification models, SVM is giving best result for the said datasets. It is also found that MWMOTE sampling is giving better result than SMOTE.

Table 2 Performance of classifiers without sampling

Dataset name	Classification model	Precision	Recall
Wineequality	SVM	0	0
	RBF	0	0
	MLP	0	0
Poker	SVM	0	0
	RBF	0	0
	MLP	0.2	0.5
kddcup- rootkit- imap	SVM	1	1
	RBF	0	0
	MLP	1	1

Table 3 Performance using SVM with sampling

Dataset name	Sampling method	Recall	Precision	Fmeasure	Gmean
Wineequality	SMOTE	1	0.75	0.8571	0.9928
	MWMOTE	1	1	1	1
Poker	SMOTE	1	0.5789	0.7333	0.9864
	MWMOTE	1	1	1	1
kddcup-rootkit-imap	SMOTE	1	1	1	1
	MWMOTE	1	1	1	1

Table 4 Performance using RBF with sampling

Dataset name	Sampling method	Recall	Precision	Fmeasure	Gmean
Wineequality	SMOTE	0.5	0.4	0.4444	0.6996
	MWMOTE	0.8750	0.8750	0.8750	0.9320
Poker	SMOTE	0.7647	0.7222	0.7428	0.8669
	MWMOTE	0.9047	0.95	0.9268	0.9495
kddcup-rootkit-imap	SMOTE	0.6666	1	0.8	0.8165
	MWMOTE	0.7948	1	0.8857	0.8915

Table 5 Performance using MLP with sampling

Dataset name	Sampling method	Recall	Precision	Fmeasure	Gmean
Wineequality	SMOTE	0.5	0.4	0.4444	0.6917
	MWMOTE	0.8	0.9231	0.8571	0.8910
Poker	SMOTE	0.7222	0.6190	0.6666	0.8380
	MWMOTE	1	1	1	1
kddcup-rootkit-imap	SMOTE	1	1	1	1
	MWMOTE	1	1	1	1

6 Conclusion

Handling imbalanced datasets is a real challenge in most of the real-life applications. Here, data-level solutions for imbalanced datasets have been described. From the simulation work, two oversampling techniques using three different classifiers SVM, RBF, and MLP are analyzed. It is found that the performance of MWMOTE is surpassing SMOTE in most of the cases. The comparative performance study of various classifiers has been shown which can be helpful for further research.

7 Future Work

Intensive research study has been made since one decade to handle imbalanced datasets. Here, as for future study, the following points can be considered:

- The overfitting problem with oversampling needs to be focused, i.e., the models need to be trained more about the generalized situations.

- Different undersampling techniques can be verified on the same data sets with the same classification models.
- Hybridization of these oversampling techniques with undersampling can be verified using the same classification models.
- Other sampling techniques like Smoteboost, Adasyn, and RAMO can be compared using the same classification model.
- These data-level solutions can be combined with algorithmic solution and checked if it is performing better.

References

1. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority oversampling technique. In: Foundations and Trends in Information Retrieva, vol. 16, pp. 321–357 (2002)
2. Chawla, N V., Lazarevic, A., Hall, O.: SMOTE Boost improving prediction of the minority class in boosting. In: The 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, pp. 1322–1328. Springer (2003)
3. Hu, S., Liang, Y., Ma, L., He, Y.: Improving classification performance when training data is imbalanced. IEEE (2005)
4. Maciejewski, T., Stefanowski, J.: Local neighborhood extension of SMOTE for mining imbalanced data. In: IEEE Symposium on Computational Intelligence and Data Mining (CIDM), pp. 978-1-4244-99 (2011)
5. Han, H., Wang, W.Y., Mao, B.H.: Borderline-SMOTE: a new oversampling method in imbalanced data sets learning. In: Proceedings International Conference Intelligent Computing, pp. 878–887 (2005)
6. He, H., Bai, Y., Garcia, E.A., Li, S.: ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: Proceedings of International Joint Conference Neural Networks, pp. 1322–1328 (2008)
7. Barua, S., Islam, M.M., Yao, X., Murase, K.: MWMOTE majority weighted minority oversampling technique for imbalanced data set learning. IEEE Trans. Knowl. Data Eng. **26**(2) (2014)
8. Jayashree, S., Alice Gavya, A.: Classification of imbalanced problem by MWMOTE and SSO. IJMTES **2**(5) (2015)
9. Japkowicz, N., Stephen, S.: The class imbalance problem: a systematic study. Intell. Data Anal. **6**(5), 429–449 (2002)
10. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. **21**(9), 1263–1284 (2009)
11. Buckland, M., Gey, A.: The relationship between recall and precision. J. Am. Soc. Inf. Sci. **45**(1), 12–19 (1994)
12. Ramentol, E., Caballero, Y., Bello, R., Herrera, F.: SMOTE-RSB: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory. Knowledge and Information System, vol. 33(2), pp. 245–265. Springer (2012)
13. Han, H., Wang, W.Y., Mao, B.H.: Borderline-SMOTE a new oversampling method in data sets learning. In: Proceedings of International Conference on Intelligent Computing, pp. 878-887 (2005)
14. Tang, Y., Zhang, Y.Q., Chawla, N.V., Krasser, S.: Modeling for highly imbalanced classification. J. latex class files. **1**(11) (2002)
15. Imam, T., Ting, K.M., Kamruzzaman, J.: z-SVM: An SVM for Improved Classification Of Imbalanced Data. Advances in Artifical Intelligence, vol. 4304, pp. 264–273 (2006)

16. Prez-Godoy, M.D., Rivera, A.J., Carmona, C.J., delJesus, M.J.: Training algorithms for radial basis function networks to tackle learning processes with imbalanced data-sets. Appl. Soft Comput. **25**, 26–39 (2014)
17. Haddad, L., Morris, C W., Boddy, L.: Training radial basis function neural networks: effects of training set size and imbalanced training sets. J. Microbiol. Methods **43**(1), 33–44 (2000)

Blended 3D Interaction Using Wii-Remote for Learning Educational Content

Suman Deb, Mitali Sinha, Sonia Nandi and Paritosh Bhattacharya

Abstract The application proposed in this paper provides an environment for physical interaction with three-dimensional models in the digital world. The main focus of the project is to introduce education to learners using modern technologies like computers and video games. An educational game is introduced, wherein the basic idea of electric charge and force has been developed. The interaction is made more physical by using a low-cost interactive device called Wii-remote which is basically a remote for a gaming console. The application effectively makes use of the motion-sensing capability of Wii-remote, thus exploiting it as an educational tool.

Keywords Three-dimensional interaction · Wii-remote

1 Introduction

With the advancement of technology, nowadays the educational domain is also shifted from traditional teaching and learning process to a modern approach of learning. The introduction of computers and video games has opened up a new

S. Deb (✉) · M. Sinha · S. Nandi · P. Bhattacharya
Computer Science and Engineering Department, NIT Agartala, Jirania, India
e-mail: sumandebcs@gmail.com

M. Sinha
e-mail: mitalisinha93@gmail.com

S. Nandi
e-mail: sonianandi90@gmail.com

P. Bhattacharya
e-mail: pari76@rediffmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_3

world for exploring innovative ways of learning [1–4]. Specific group of learners which comprises children of age below 11–14 shows more interest in learning when a game-based approach is cultivated in the field of education [5, 6].

Various projects have been developed in the area of edutainment which embeds technology into education [7, 8]. Unlike most of the video game applications that provide a two-dimensional experience to the learners, our application will be able to provide a three-dimensional environment where the user can physically manipulate objects in the digital world. A very low-cost interactive device called Wii-remote has been used as a graspable tool and it has also been used by few researchers for experimenting in various domains [9–12]. The Wii-remote is basically a remote for a gaming console called Wii, which contains built-in sensors for gesture recognition and position detection. The sensors of the Wii-remote have been exploited to introduce communication between the physical Wii-remote and the digital objects in the digital world. The gesture recognition ability of Wii-remote is mapped to the objects in the digital world. For a three-dimensional approach, the application is developed in 3D modeling software called blender and runs in unity game engine. Thus by exploiting the sensory abilities of game controllers, we can develop educational software for providing a better environment for learning where children's interest and engagement are accelerated.

2 Proposed System

2.1 Component Structure

In 2006, a Japanese company Nintendo Co. Ltd. released a gaming console named Wii which gave a good competition to the various powerful gaming consoles that already existed like Microsoft's Xbox 360 and the Sony PlayStation. The remote developed for the gaming console named Wii-remote became the center of attraction very soon. The Wii-remote arrived with the sensors built in it which made it a powerful tool providing motion detection as well as gesture recognition rather than using it as a simple pointing device. A built-in accelerometer of the Wii-remote is used to track the movement and the infrared sensor at the tip of the Wii-remote is used to detect its position. The Wii-remote is connected to the computer using simple Bluetooth connection mechanism, and as a result, it provides a very easy communication service. The data from the Wii-remote is transmitted to the computer via Bluetooth connection and then the data is manipulated to work within developing the application. The infrared camera requires an infrared light source for properly communicating the data. Two infrared boxes containing infrared lights are developed from which the rays are captured by the Wii-remote and then it communicates with the computer via Bluetooth (Fig. 1).

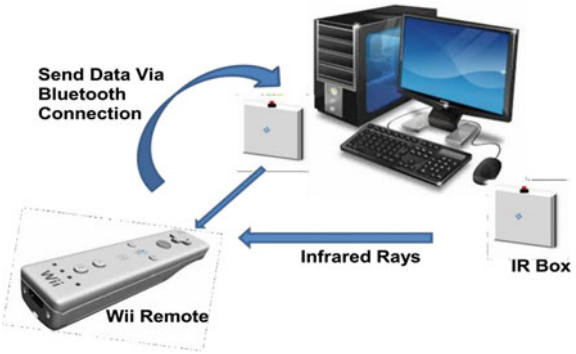


Fig. 1 Data transfer mechanism of Wii-remote

2.2 Application Design

The application is designed in such a way so as to give a better user experience in learning. Special attention is provided on building the interface keeping in mind different factors like color combinations and various audiovisual effects. Different modules of the system development are shown in Fig. 2. The application is developed in three-dimensional modeling software called blender and the unity 3D game engine. The application is based on teaching the concept of science related to electric charges and force. A cube is created that will carry positive or negative electric charge which will be defined by the player at first. The cube is mapped with the Wii-remote and thus it can be manipulated using the Wii-remote in the digital world. The scene will also contain some suspended game objects that will resemble positive or negative charges denoted by plus and minus signs, respectively. As the player brings the cube closer to the suspended charged particles in the scenes, the

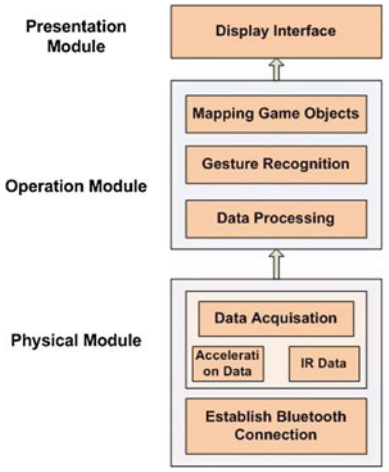


Fig. 2 System interconnection module

particles attract or repel according to the laws of attraction and repulsion, that is, if both the objects attain equal charges, then they will attract; otherwise they will repel. The game is developed in c# language and the interface is designed in unity 3D.

3 Results and Discussion

The evaluation of the system was carried out at two levels. First, the effectiveness and accuracy of the system in selecting and dragging a game object were tested. Next, the system was evaluated to check whether it is able to increase student's motivation in the learning process.

3.1 Performance Evaluation of the System

The system evaluation was carried out with the assistance of 10 subjects. The subjects comprised students in the age group 12–14, with basic computer knowledge and a little experience with gaming consoles. Experiments were conducted with the 10 subjects, where they were asked to select the game objects (resembling charged particles) suspended in the virtual world on screen by pressing the Wii-remote's A button. Then, drag the selected game object toward a target set in the virtual world. Each subject performed the experiment at several attempts and a number of attempts succeeded and failed were recorded. Table 1 reports the accuracy measures of the proposed system.

Table 1 System's performance evaluation

Subjects	Success		Failure		Accuracy (%)
	TP	TN	FP	FN	
I	7	2	0	1	87
II	9	0	1	0	90
III	9	0	1	0	90
IV	8	1	0	1	88
V	7	2	0	1	87
VI	9	0	1	0	90
VII	8	1	1	0	88
VIII	9	0	0	1	90
IX	8	1	0	1	88
X	9	0	1	0	90
			Average Accuracy		88.80

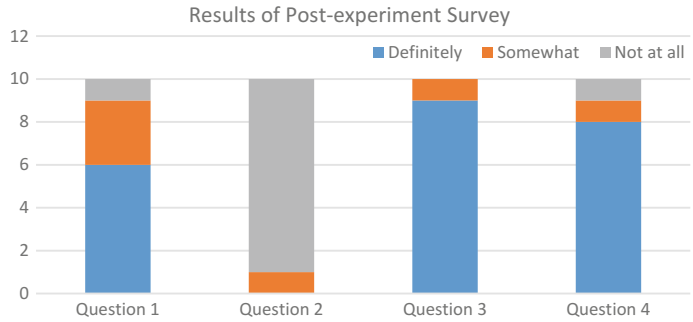


Fig. 3 Results of post-experiment survey

3.2 Arousing Interest Among the Subjects

The system was evaluated to check whether it is able to accelerate student's interest and motivate them in the learning process. After completing the experiment, the subjects were presented with questionnaires for investigating the usability and ease-of-use of the proposed system for interacting with the objects in the virtual world. The result of the post-experiment survey and their results are shown in Fig. 3. The post-experiment survey results gave evidence of the feasibility and effectiveness of the proposed system. It shows that the system is interesting and is able to accelerate the learning process with increased motivation among the subjects. Another motive of the survey was to investigate whether the system provided them an easy way of learning different educational contents, which showed positive results as shown in Fig. 2. The subjects also showed interest in recommending the system to other people as they found it both educational and playful ways of learning new concepts.

Question 1 Are you more interested in learning new concepts at the end of this course?
Question 2 Do you think the concepts were difficult to learn and understand?
Question 3 Do you recommend your friends to try our system?
Question 4 Do you find the learning interesting and enjoying?

4 Conclusion

Our goal was to develop a system that can retain the interest of young learners by providing them education along with entertainment. Learners are provided an interactive environment to explore some concepts of science in a playful manner. The learners will also get feedbacks from each of their actions which they will perform using the Wii-remote controller in the form of visual and auditory

feedbacks. The post-experiment surveys were conducted to investigate the effectiveness of the proposed system in motivating the young learners. The survey results displayed that the system was successful to increase student's interest in learning by blending technology and educational content in an interesting manner.

Declaration Authors have obtained all ethical approvals from appropriate ethical committee and approval from the students or from their parents/LAR (because the students are minor) who participated in this study.

References

1. Papastergiou, Marina: Digital game-based learning in high school computer science education: impact on educational effectiveness and student motivation. Comput. Educ. **52** (1), 1–12 (2009)
2. Shih, K., Squire, K., Lau, R.: Guest editorial: special section on game-based learning. IEEE Trans. Learn. Technol. **3**(4), 278–280 (2010)
3. Quing, L., Tay, R., Louis, R.: Designing digital games to teach road safety: a study of graduate students' experience. Loading **6**(9) (2012)
4. Virvou, M., Katsionis, G., Manos, K.: Combining software games with education:evaluation of its educational effectiveness. Educ. Technol. Soc. **3**(4), 307–318 (2005)
5. Annetta, L.: The "I's" have it: a framework for serious educational game design. Rev. Gen. Psychol., 105–112 (2010)
6. Durkin, k.: Video games and young people with developmental disorders. Rev. Gen. Psychol., 122–140 (2010)
7. Barah, S., et al.: Making learning fun: Quest Atlantis, a game without guns. Educ. Technol. Soc. **53**, 86–107 (2005)
8. Jong, M., et al.: An evaluative study on VISOLE—virtual interactive student-oriented learning environment. IEEE Trans. Learn. Technol. **3**, 307–318 (2010)
9. Bryant, J., Akerman, A., Drell, J.: Wee Wii: preschoolers and motion-based game play. Annual meeting of the International Communication Association, Montreal, Quebec, Canada (2008)
10. Ho, J.H., et al.: Investigating the effects of educational game with Wii remote on outcomes of learning. Transactions on Edutainment III, pp. 240–252. Springer Berlin Heidelberg (2009)
11. Hwang, J.Y., Yi-Luen, E.: WiiInteract: designing immersive and interactive application with a Wii remote controller. In: 15th International Conference on Computer Games. AI, Animation, Mobile, Interactive Multimedia, Educational & Serious Games, July 28–31
12. Holmquist, L.E., Ju, W., Jonsson, M., Tholander, J., Ahmet, Z., Sumon, S.I., Acholonu, U., Wingrad, T.: Wii sciene:teaching the laws of nature with physically engaging video game technologies, Stanford University

Augmented Use of Depth Vision for Interactive Applications

Sonia Nandi, Suman Deb and Mitali Sinha

Abstract In this paper, an interactive system based on hand gesture recognition is proposed. We devised a way to use signed communication between electric appliances. Our system is inspired by different hand gestures that we use to make someone understand our mind. The proposed prototype will make the electric appliances understand the hand gestures made by the user in front of Kinect. The system will be able to detect the specified hand gestures made by the user and also able to recognize the meaning of the gestures, and depending on that the electric appliances will react. This interactive system can be integrated into a variety of application in daily life.

Keywords Kinect · Sign language · Depth perception · Depth vision
Interactive application · Hand gesture recognition · Hand pose

1 Introduction

Hand gesture is a mode of communication mainly for hearing challenged people. But sometimes this silent communication is also needed by other people depending on the situation. Hand gestures can be recognized using video camera and wearable sensors, for example, data gloves. In this paper, we tried to find a natural and easy way using which we can interact with different electrical appliances by our hand gestures. To provide such interactivity, depth sensing can be effectively used.

Depth perception is the ability to judge the position of an object whether it is nearer or far away from the objects. It also helps us to form an informal idea about

S. Nandi (✉) · S. Deb · M. Sinha
National Institute of Technology, Agartala, India
e-mail: sonianandi90@gmail.com

S. Deb
e-mail: sumandebcs@gmail.com

M. Sinha
e-mail: mitalisinha93@gmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_4

the speed of the object coming toward us. In this paper, it is tried to find out different possibilities of depth sensor in interactive applications.

Modern researchers have provided a variety of depth-sensing devices having features like object detection, position detection, motion and skeletal tracking, etc.; these depth-sensing devices are playing an important role in medical, robotic, learning technologies, and many other fields.

There are various interactive devices other than keyboard and mouse, the so-called Natural User Interaction (NUI) [1]. These devices do not require any special medium while interacting. Nowadays, a variety of input channels are available, for example, speech-based control, eye movement-based control, etc. Among all the interaction techniques, gestural-based body tracking movement using computer vision is getting popular. This paper presents an idea on the use of depth for hand gestures to create user interfaces.

2 Literature Review

This section focuses on the research of various interactive devices using human tracking mainly in our home environment. Smart fan [2] developed will track the presence of the person in the room using ultrasonic sensor and depending on the room's temperature, the fan's speed will get controlled. But the limitation with ultrasonic sensor is that it cannot differentiate between human and any other obstacle present in front of it.

This section reviews the previous research done on different hand gesture recognition devices. GestureWrist and GesturePad [3] are used together to recognize hand and forearm movements. KHU-1, a gesture recognition system using data gloves, is developed for 3D hand tracking [4], which uses PC and a Bluetooth device. Another visual motion data glove [5] is developed which uses a single-channel video to track hand movements, but all these wearable devices are a bit uncomfortable to use. So, depth sensor Kinect can be a good alternative for designing interactive devices.

3 Depth Sensor—Kinect

Depth sensor Kinect consists of IR camera and IR laser projector. The raw depth data can be extracted by the principle of structured light. The IR emitter emits IR pattern in the screen. These IR patterns get detected by the IR camera. These patterns help the processor compute the IR image. Microsoft Kinect can track up to six users' presence in front of the viewing angle of the sensor along with the skeletal data of two skeletons. It can also track twenty joints of a person. But Kinect SDK cannot detect the hand joints by itself. Special programming is needed to make Kinect recognize hand gestures. In our solution, we have used k-curvature algorithm to detect hand gestures (Figs. 1 and 2).

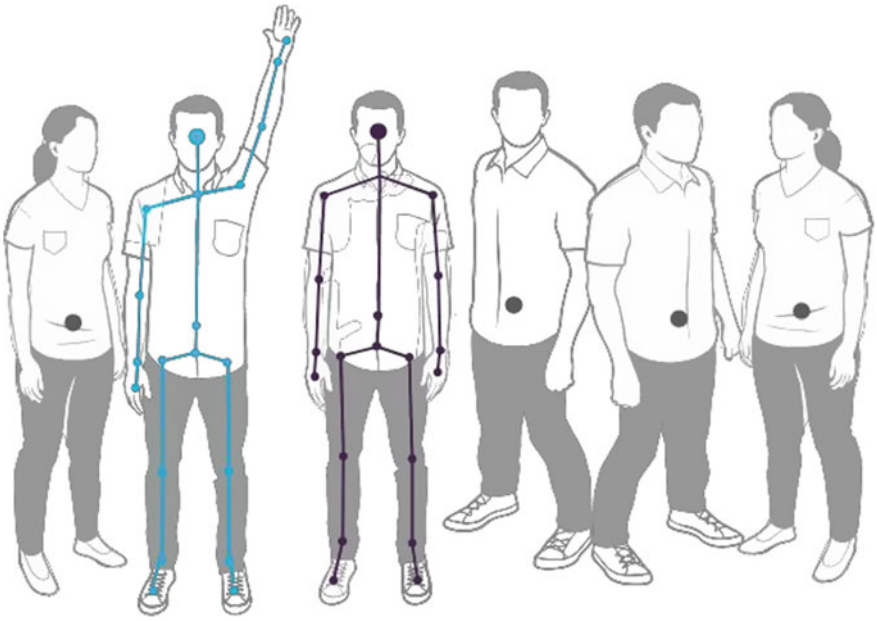


Fig. 1 Kinect recognizing six persons at a time [6]

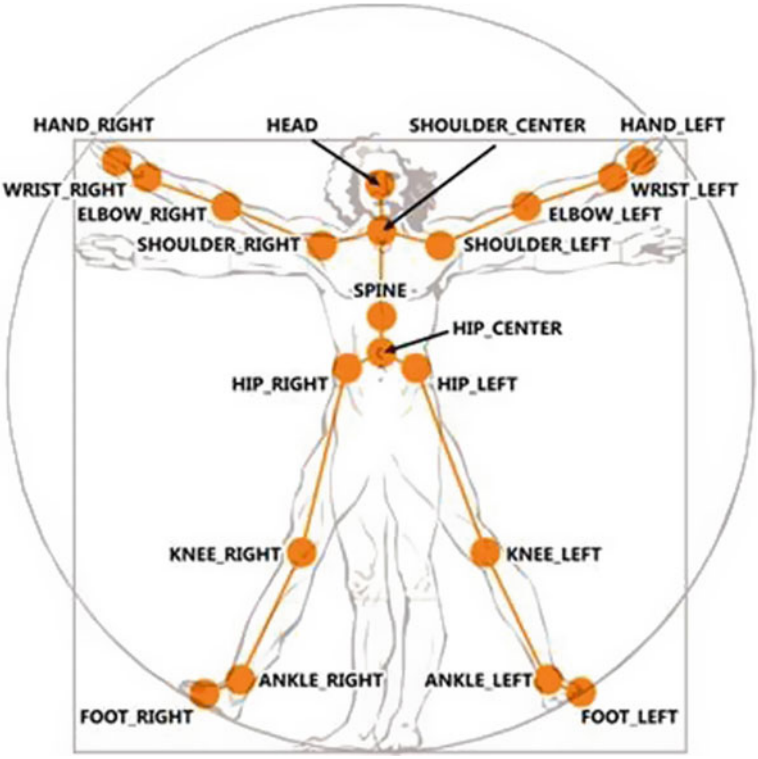


Fig. 2 Kinect tracking 20 joints [7]

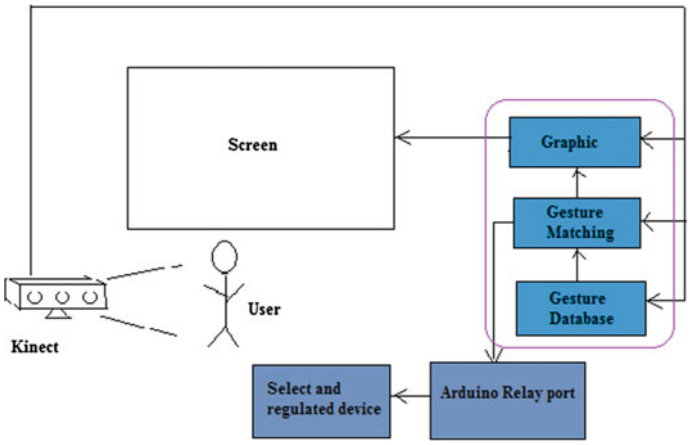


Fig. 3 System architecture

4 System Architecture

The architecture of the system includes five components: graphics, gesture matching, gesture database, Kinect as gesture tracker, and Arduino relay port. Figure 3 shows the relationship between all the five components. User's hand, finger, and skeletal movement are tracked using Kinect and then the raw gesture data is compared with the gesture database through the gesture matching component. This gesture matching component sends the data to relay port to select and regulate device accordingly. The graphics component visualizes the hand and finger movements by the user.

5 Methodology

The ideology represented in the paper is to make household electric appliances interactive. Depending on the elbow movement, i.e., the angle made by the shoulder and the elbow joint, the user can select different appliances [8] and by making different hand gestures, the user can control or regulate the selected appliance.

Fig. 4 Rotating fingers clockwise direction

Fig. 5 Rotating anticlockwise direction

6 Representation

In the figure below, it is described how the gradient feeling can be introduced by making different angles and figure gestures with the help of Kinect. We can control different electric devices by rotating our fingers clockwise as shown in Fig. 4; we can increase the volume of T.V suppose and by rotating anticlockwise as shown in Fig. 5, we can decrease the volume (Fig. 6). Also, we can start the selected device

Fig. 6 Rotating palm right to left

Fig. 7 Rotating palm upside down

by moving our palm from right to left as shown in Fig. 7 and stop the device by moving palm upside down as shown in Fig. 7.

7 Conclusion

Proposed interactive system can help the elderly and disabled person to operate different appliances easily. Our main target is to design a smart home environment. The designed platform will track the person and help him to select the appliance and regulate according to his need. This interactive system is capable of providing great comfort to the user (Fig. 8).

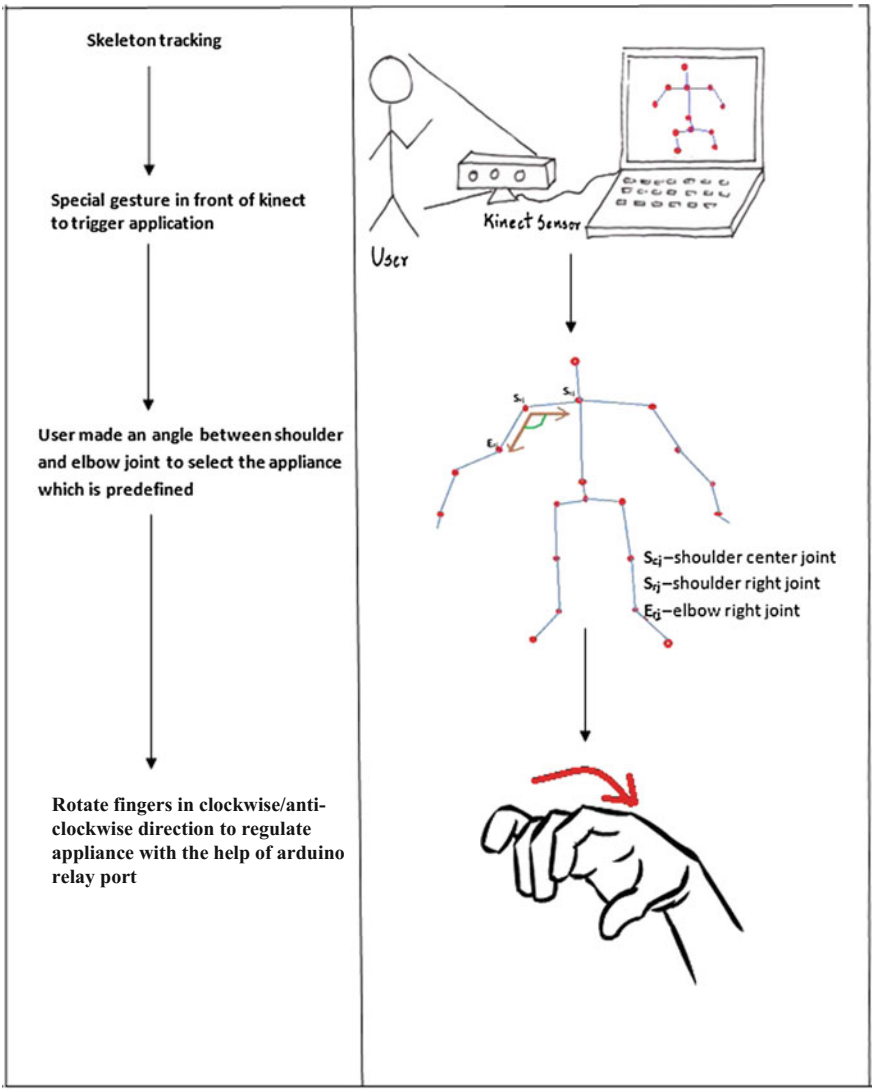


Fig. 8 Schematic of the methodology

References

1. Steinkirch, M.V.: On depth sensors applications for computational photography and vision, State University of New York at Stony Brook, May 20, 2013
2. Ishrat, T., Rahaman, M.A., Ahammad, A.: Smart fan for human tracking. In: 2014 9th International Forum on Strategic Technology (IFOST), 21 Oct 2014, pp. 124–127. IEEE

3. Rekimoto, J. Gesturewrist and gesturepad: Unobtrusive wearable interaction devices. In: Fifth International Symposium on Wearable Computers, 2001, pp. 21–27. Proceedings. IEEE (2001)
4. Kim, J.H., Thang, N.D., Kim, T.S.: 3-d hand motion tracking and gesture recognition using a data glove. In: 2009 IEEE International Symposium on Industrial Electronics, 5 July 2009, pp. 1013–1018. IEEE
5. Han, Y.: A low-cost visual motion data glove as an input device to interpret human hand gestures. IEEE Trans. Consumer Electron. **56**(2), 501–509 (2010)
6. https://msdn.microsoft.com/en-us/library/hh973074.aspx
7. http://www.contentmaster.com/kinect/kinect-sdk-skeleton-tracking/
8. Nandi, S., Deb, S., Sinha, M.: Creating low cost multi-gesture device control by using depth sensing. In: Information Systems Design and Intelligent Applications 2016, pp. 105–113. Springer, India
9. Li, Y.: Hand gesture recognition using Kinect. In: 2012 IEEE International Conference on Computer Science and Automation Engineering, 22 Jun 2012, pp. 196–199. IEEE
10. Chan, M., Estève, D., Escriba, C., Campo, E.: A review of smart homes—present state and future challenges. Comput. Methods Prog. Biomed. **91**(1), 55–81 (2008)

An Enhanced Intrusion Detection System Based on Clustering

Samarjeet Borah, Ranjit Panigrahi and Anindita Chakraborty

Abstract The aim of a typical intrusion detection framework is to recognize attacks with a high discovery rate and low false alarm rate. Many algorithms have been proposed for detecting intrusions using various soft computing approaches such as self-organizing map (SOM), clustering etc. In this paper, an effort has been made to enhance the intrusion detection algorithm proposed by Nadya et al. The proposed enhancement of the algorithm is done by adding the SOM training process. Clustering of the data is done to differentiate abnormal data from the normal data. The clustered data may sometime contain both normal and abnormal data thus leading to false alarms. In this regard, k-means algorithm is further used to detect those abnormal data and reducing the rate of false positive. The SOM is trained using the neural network toolbox present in Matlab R2010b. The enhanced algorithm yields desired results both in terms of higher detection rates and removal of false positives.

Keywords Intrusion detection · Attack · Clustering · MatLab
False positive · False negative · SOM

1 Introduction

The action of intrusions gradually weakens the confidentiality of resources and information present or to hamper the integrity and availability of behavior in a host or in a network environment. Therefore, intrusions are any sets of activities threatening the veracity, confidentiality, or accessibility of a network resource [1–3]. It can also provide unofficial access to important and useful information and unauthorized file modification which are reasons behind harmful activities.

S. Borah (✉) · R. Panigrahi · A. Chakraborty
Sikkim Manipal Institute of Technology, Sikkim Manipal University,
Rangpo, Sikkim, India
e-mail: samarjeetborah@gmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_5

To counter such malicious activities, an intrusion detection system (IDS) comes into action. The IDS identifies these unauthorized activities and takes appropriate action, thus preventing them at real time [4–8]. It is used for detecting real-time monitoring system activities and real-time aggressive behavior, and takes corrective measure to avoid or minimize the occurrence of attacks. The IDS monitors the events that are occurring in the system or networks and analyzes them for intrusion. It collects information related to events taken place in a system and trigger an alarm, when an intrusion is detected, thus preserving the data integrity from attacks. It also helps in handling and monitoring of audit trails, assessment of their system, and networks which is an important part of security management.

This paper discusses the refinement of an existing algorithm [9] which can be used for host-based intrusion detection. In this research work, KDD99 cup dataset is used which consists of both normal and abnormal data and so we first use the k-means clustering algorithm to differentiate the normal from the abnormal data and then train the SOM according to the dataset.

2 Motivation

This work is motivated by the work of El Moussaid et al. [9]. They proposed an improved k-means clustering algorithm. They have tested the algorithm with KDD'99 dataset. Four different types of attacks were identified such as DOS, U2R, R2L, and Probe. In their approach, out of the 41 features of the dataset, 13 features were selected reducing the noise, dataset dimension, and time complexity. Normalization of dataset is done by calculating the mean absolute deviation and standardized measured. The normalized data is then subjected for cluster number initialization. This is done by calculating the similarity and clustering them in one group, while clustering if records remained to be cluster is less than ten percentages of total records is clustered in the same group. This may sometimes lead to a combined cluster of both normal and abnormal data resulting in large number of false-positive alarms. To reduce these alarms, k-means clustering is used by calculating the Euclidean distance and density of each connection record. They have labeled the clusters by calculating the percentage of abnormal connection "ɵ". If the member of cluster is less than or equal to the product of ɵ, then total records N are labeled as anomaly otherwise as normal. It has been found that the algorithm gives better results for DOS and R2L attacks and the false-positive rate is 30%.

The main advantage of the approach is that it is less time consuming compared to other such approaches. But for Probe and U2R attacks, the rate of detection is low as compared to other attacks. The proposed enhancement of the algorithm is done by adding the SOM training. It tries to reduce the false-positive alarm rate by using k-means clustering.

3 Working Methodology

In the proposed enhancement, the approach consists of several modules, each for different purposes. The modules are described below:

3.1 Dimensionality Reduction

The dataset considered is subjected to dimensionality reduction which is done using the Principal Component Analysis (PCA) (Fig. 1).

3.2 Clustering

The initial clustering is performed as per the existing method. For each feature of the dataset taken, the minimum and maximum values are found out, and next it helps in finding the upper and lower limits of every feature; if each value of the features has a value between lower and upper limits, then they are grouped into one cluster. Here, one condition is observed; if the amount of connection records is fewer than 10% of the total records of the dataset, then they are grouped into same cluster. In this module, we are trying to define the number of cluster before it is subjected to k-means algorithm. This helps us in getting an idea as to how many clusters a particular dataset may contain. Thus, it may be possible that a particular cluster or more may contain the intrusion information and it may not be necessary to check all the clusters available. The Euclidian distance for each feature is calculated for every connection record. The k-means algorithm is used for the clustering process. The initial centroids are assumed randomly. Distance between the centroid and all the data points are calculated, and the minimum distance is considered resulting as

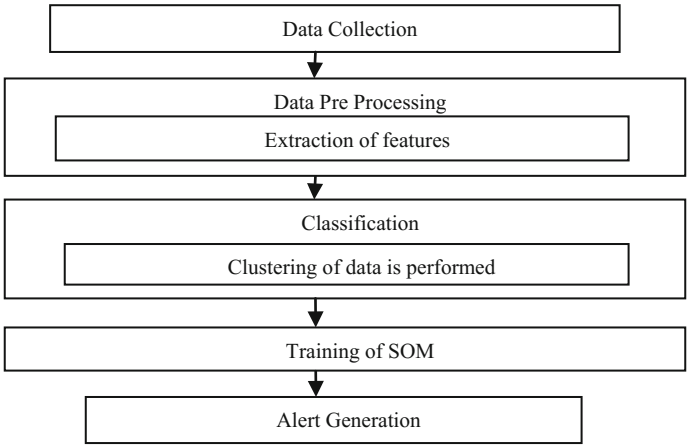


Fig. 1 Architectural diagram

the next centroid leading to cluster updation. This process continues till a maximum repetition is reached defined beforehand [10].

3.3 Training of Self-organizing Map

Self-organizing map is trained using batch algorithm in neural network toolbox of Matlab R2010b. The training runs for maximum of 200 epochs. During training, the weight vector associated with each neuron is moved to the center of the cluster, thus reducing high dimensionality input into two dimensions of topology. This is quite a beneficial tool because after training of SOM there are various options which help us in visualizing the resulting clusters. The various options available are SOM topology, SOM neighbor connections, SOM neighbor distance, SOM weights, SOM sample hits, and SOM weight positions.

4 The Proposed Algorithms

4.1 Data Clustering

```
START
Calculate Max (fi) and Min (fi) of each feature fi.
Calculate upper limit (UL) and lower limit (LL) of
features (fi) as:
    UL= [(Max+Min)/2]
LL= [(Max-Min)/2]
For each feature fi
    if value of LL ≤  fi  ≥ UL
          Insert 1 in table
    Else
          Insert 0 in table
    End if
    For k = 1 to max_row
          For m = k+1 to max_row
                Compare (k, m)
                if similar
                      Keep in same cluster
                Else
                      Keep in different cluster
                End if
          End for
    End for
    If records remained < 10% of total records
          Keep in same cluster
    Else
         Repeat from line 7
    End if
End for
STOP
```

4.2 False-Positive Reduction

```
START
Input_Data= mixed data obtained from clustering of data
in 5.1
Apply K-means clustering algorithm to input data
This step continues till MaxRepeat (defined beforehand)
STOP
```

4.3 Self Organizing Map (SOM) Training

The training phase is having the following steps:

(a) Apply Principal Component Analysis (PCA) for dimentionality reduction.
(b) Consider the score with highest significance.
(c) Train SOM using the batch algorithm.

5 Implementation

The refined algorithm is implemented in Matlab on windows platform. The detection process involves several steps. The preprocessing step is performed prior to the cluster number initialization. The SOM is trained using neural network toolbox. Preprocessing reduces the dimensions of data. The batch algorithm is used for SOM training. Here, the whole training set is trained only once and after that the map is updated having the net effect of the samples. The default training algorithm for SOM is batch algorithm as it is very much faster to calculate than the normal sequential algorithm (in Matlab).

Dataset used for this experiment is KDD'99 cup dataset. A nice description of the same is found in [11, 12]. It has a set of 41 features out of which there are few nominal features. Thus, in order for appropriate clustering, the nominal features are converted into numeric values as follows (Table 1).

Table 1 Transformations table

Name	Value
UDP Protocol	1
ICMP Protocol	2
TCP Protocol	3
Flags	4–16
Services	7–15

6 Results

While comparing, two points are considered. If the remaining records are less than 10% of the total records, they are grouped into the same cluster; the rest of the records were grouped into cluster which results in false-positive alarm generation.

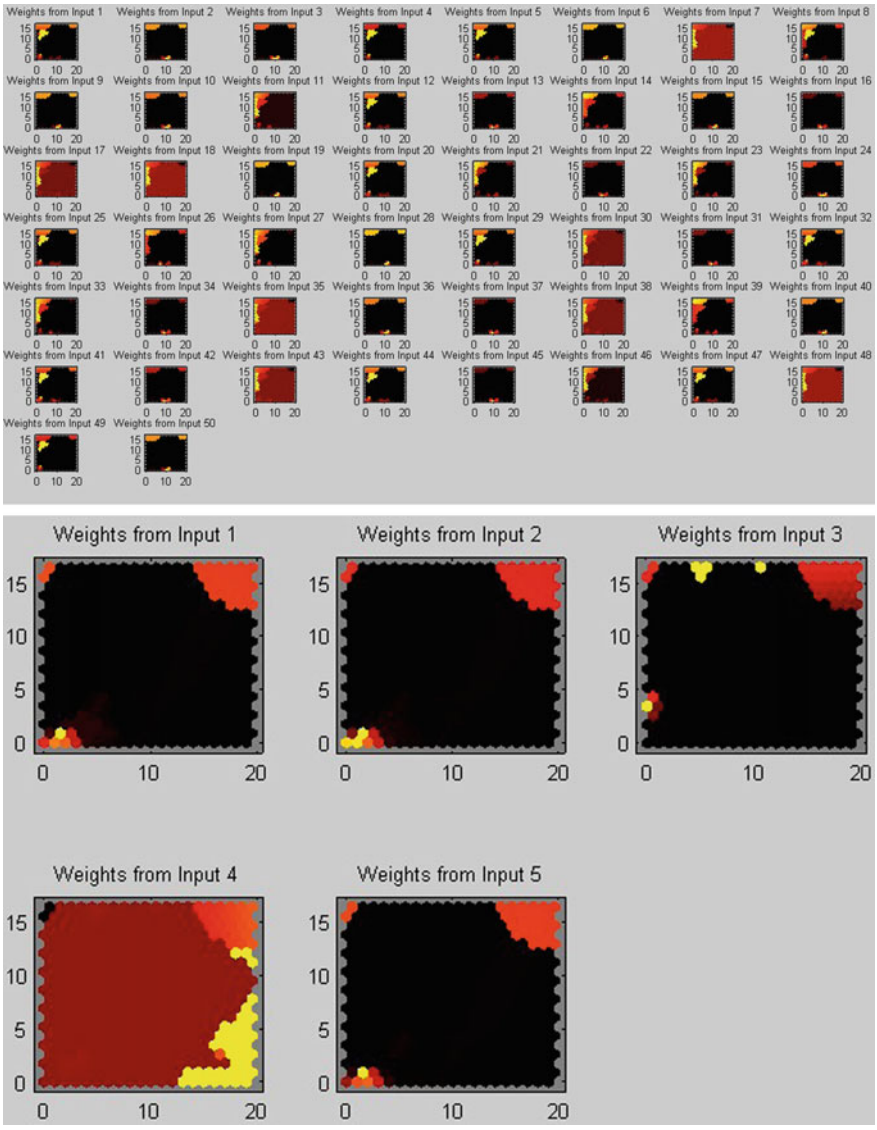


Fig. 2 SOM input planes

K-means clustering is applied to cluster and thus the abnormal data is differentiated from the normal data by clustering the normal data in cluster3 and the abnormal data in cluster1. But still some normal data is still considered as abnormal data, and as in record 4, is normal data but is considered as abnormal data.

The training of the SOM is done using the batch processing algorithm in neural network toolbox.

The SOM input planes show for each input features a weight plane. They are weight visualizations which connect the every input with the each of the neurons. The samples having dark color represent large weight and the light color represent smaller weight. Any two or more samples having same weight may have their color be same. The SOM input plane for the clustering of data and re-clustering using k-means is shown in Fig. 2.

To evaluate the detection rate for host-based intrusion, the dataset considered is clustered by the k-means and is trained using self-organizing map. The algorithm was tested with both the labeled and unlabeled data.

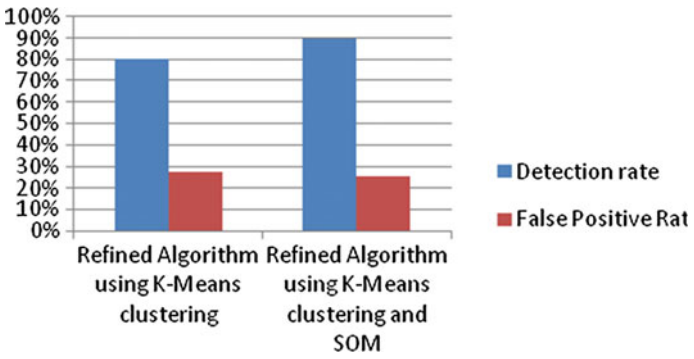


Fig. 3 Detection rate for refined algorithm

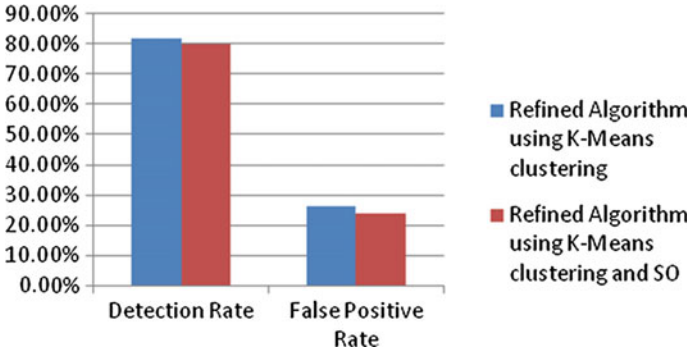


Fig. 4 False-positive rate for refined algorithm

Table 2 Results found for labeled data

For labeled data	Detection rate (%)	False alarm rate (%)
Refined algorithm using K-means clustering	80	27.33
Refined algorithm using K-means clustering and SOM	90	25

Table 3 Results found for unlabelled data

For labeled data	Detection rate (%)	False-positive rate (%)
Refined algorithm using K-means clustering	81.75	26
Refined algorithm using K-means clustering and SOM	80	24

Following figures show the detection rate of the host-based intrusion while being clustered by k-means clustering and trained using labeled data and unlabelled data. Thus, we try to reduce the false-positive alarm rate using this process as shown in figure.

The comparative performance analysis is shown in Figs. 3 and 4, and Tables 2 and 3.

7 Conclusion

The initialization of cluster in the beginning helps us in working with a finite number of clusters. K-means clustering helps in fast computing since the number of clusters is defined beforehand. However, the initialization plays an important role as different values will give different results. The k-means clustering algorithm is able to reduce the percent of false-positive alarm rate.

In this paper, detection rates for k-means clustering with and without SOM training are shown for refined algorithm. The approach leads to the conclusion that when SOM is trained for detecting unlabelled intrusions, it may or may not be able to generate good results. Thus, the SOM must be trained repeatedly using different numbers of nodes until the improved results are found. The false-positive alarm rates can be increased by clustering the data efficiently so that it is able to differentiate correctly between the normal data and abnormal data and the training of SOM must be done accordingly.

References

1. Luo, N., Yuan, F., Zuo, W., He, F., Zhou, Z.: Improved unsupervised anomaly detection algorithm. In: Proceedings of Third International Conference, RSKT 2008, Chengdu, China, 17–19 May 2008. Springer Rough Sets and Knowledge Technology Series (2008)
2. Youssef, A., Emam, A.: Network intrusion detection using data mining and network behaviour analysis. Int. J. Comput. Sci. Inf. Technol. (IJCSIT) **3**(6), 87–98 (2011)
3. Suryavanshi, M., Akiwate, B., Gurav, M.: GNP-based fuzzy class-association rule mining in IDS. Int. J. Emerg. Trends Technol. Comput. Sci. (IJETTCS) **2**(6), 179–183 (2013). ISSN 2278-6856
4. Beal, V.: Intrusion Detection (IDS) and Prevention (IPS) Systems. http://www.webopedia.com/DidYouKnow/Computer_Science/intrusion_detection_prevention.asp (2005). Accessed 15 July 2005
5. Kazienko, P., Dorosz, P.: Intrusion Detection Systems (IDS) Part I—(network intrusions; attack symptoms; IDS tasks; and IDS architecture). http://www.systemcomputing.org/ssm10/intrusion_detection_systems_architecture.htm (2003). Accessed 07 Apr 2003
6. Borah, S., Chakravorty, D., Chawhan, C., Saha, A.: Advanced Clustering based Intrusion Detection (ACID) Algorithm, Advances in Computing and Communications, Springer CCIS series, Vol. 192, Part 1, ISSN: 1865:0929, pp. 35–43, (2011) http://dx.doi.org/10.1007/978-3-642-22720-2_4
7. Borah, S., Chakraborty, A.: Towards the Development of an Efficient Intrusion Detection System. Int. J. Comput. Appl. **90**(8), 15–20 (2014)
8. Dutt , I., Borah, S., Maitra, I.: Intrusion Detection System using Artificial Immune System. Int. J. Comput. Appl. **144**(12),19–22 (2016)
9. El Moussaid, N., Toumanari, A., Elazhari, M.: Intrusion detection based on clustering algorithm. Int. J. Electron. Comput. Sci. Eng. **2**(3), 1059–1064. ISSN- 2277-1956
10. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, pp. 281–297. MR 0214227. Zbl 0214.46201 (1967)
11. Olusola, A.A., Oladele, A.S., Abosede, D.O.: Analysis of KDD '99 intrusion detection dataset for selection of relevance features. In: Proceedings of the World Congress on Engineering and Computer Science 2010 Volume I, WCECS 2010, 20–22 Oct 2010, San Francisco, USA (2010). ISBN: 978-988-17012-0-6, ISSN: 2078-0958
12. Kayacik, H.G., Zincir-Heywood, A.N., Heywood, M.I.: Selecting features for intrusion detection: a feature relevance analysis on KDD 99. In: Third Annual Conference on Privacy, Security and Trust (PST), 12–14 Oct 2005, The Fairmont Algonquin, St. Andrews, New Brunswick, Canada (2005)

Identification of Co-expressed microRNAs Using Rough Hypercuboid-Based Interval Type-2 Fuzzy C-Means Algorithm

Partha Garai and Pradipta Maji

Abstract MicroRNAs are a class of small RNA molecules, which play an important regulatory role for the gene expression of animals and plants. Various studies have proved that microRNAs tend to cluster on chromosomes. In this regard, a novel clustering algorithm is proposed in this paper, integrating rough hypercuboid approach and interval type-2 fuzzy *c*-means. Rough hypercuboid equivalence partition matrix is used here to compute the lower approximation and boundary region implicitly for the clusters without the need of any user-specified threshold. Interval-valued fuzzifier is used to deal with the uncertainty associated with the fuzzy clustering parameters. The effectiveness of proposed method, along with a comparison with existing clustering techniques, is demonstrated on several microRNA data sets using some widely used cluster validity indices.

Keywords Co-Expressed microRNA clustering · Rough hypercuboid · Interval type-2 fuzzy c-means

1 Introduction

MicroRNAs (miRNAs) are short, non-coding endogenous RNAs which can regulate gene expression at post-transcriptional level by directing their target mRNAs for degradation or translational repression. Recent discoveries revealed that a moderate fraction of miRNA genes is likely to form a cluster. The miRNAs forming a cluster that is in close proximity on chromosomes have a great chance to be processed as co-transcribed units. So the miRNAs in a cluster are mostly co-expressed. Expression profiling of miRNAs generates a large amount of data. Complicated gene interac-

P. Garai (✉)
Department of Computer Application,
Kalyani Government Engineering College, Kalyani, India
e-mail: parthagarai@gmail.com

P. Maji
Biomedical Imaging and Bioinformatics Lab,
Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India
e-mail: pmaji@isical.ac.in

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_6

tion networks increase the challenges of comprehending and interpreting the resulting huge amount of data. Clustering has retained high interest for analyzing miRNA data, where miRNAs are grouped into a number of clusters according to their similarities [1]. The grouping is done in such a way that the degree of association between miRNAs from the same cluster is maximum and minimum for the miRNAs from different clusters [2]. The purpose of miRNA clustering is to make a collection of co-expressed miRNAs that show cofunction and co-regulation.

The traditional hard c-means (HCM) algorithm [1] is one of the primitive and most widely used objective function-based clustering algorithms, which minimizes the sum of the distances between the miRNAs and their corresponding centroid. Each miRNA has membership in only one of the clusters. Uncertainty management is one of the important issues in miRNA data analysis. A major part of this uncertainty comes from incompleteness and vagueness in cluster definition. Fuzzy set theory, which is derived from the classical set theory, provides gradual membership values for the miRNAs, relaxing the requirement of the HCM. In this background, fuzzy c-means (FCM) algorithm is proposed, where the memberships of the miRNAs decrease as their distance from the cluster center increases. It helps to deal with the miRNA data that has membership in multiple clusters at the same time.

Rough set [3], proposed by Pawlak, is a very useful paradigm, which can be used to deal with uncertainty, vagueness, and incompleteness associated with data for the cluster definition and hence has been widely used for clustering in uncertain space. Fuzzy sets and rough sets can be combined for the use in the domain of reasoning with uncertainty [4, 5]. Integrating both rough and fuzzy sets, Maji and Pal proposed the rough-fuzzy c-means algorithm (RFCM) [6], encapsulating two related and complementary, but distinct concepts generated from uncertainty in knowledge, vagueness, related to fuzzy set, and indiscernibility concept, related to rough sets. The algorithm provides the boundary region with gradual membership values, which can group the miRNAs well in a data set with the presence of uncertainty.

Both FCM and RFCM algorithms use type-1 (T1) fuzzy sets. Although the researchers found some limitations of T1 fuzzy sets, lots of researchers are still using this theory. Zadeh proposed the generalized version of T1 fuzzy set, called type-2 (T2) fuzzy set [7]. T2 fuzzy set has fuzzy membership grades. It can be used in the situation where the actual membership function for some fuzzy set is not known. The combination of T2 fuzzy set theory and rough set theory is called the T2 fuzzy-rough sets and can be readily used where the exact membership function cannot be determined easily. T2 fuzzy sets have three-dimensional membership function, and so more degrees of freedom compared to the T1 fuzzy sets. Rhee and Hwang proposed a T2 fuzzy c-means based algorithm (T2FCM) [8] by assigning membership grades to the T1 memberships. An IT2 fuzzy c-means (IT2FCM) algorithm [9] has been proposed by Hwang and Rhee. The IT2 fuzzy concept is also successfully applied in various pattern recognition applications [10]. Hence, the combination of IT2 fuzzy sets and rough sets can be a useful technique for clustering.

In this paper, a hybrid algorithm (RH-IT2FCM) is proposed, integrating the benefits of rough hypercuboid approach, probabilistic membership of IT2 fuzzy sets, and c-means algorithm. While the concept of lower approximation and bound-

ary region of rough sets deals with uncertainty, vagueness, and incompleteness in miRNA cluster definition, the use of fuzzy membership of fuzzy sets in the boundary region enables efficient handling of overlapping partitions in uncertain environment. The cluster prototypes are dependent on the weighted average of the crisp lower approximation and probabilistic boundary. The proposed algorithm can find overlapping and vaguely defined clusters. Partitioning the miRNAs in lower approximation and boundary region is done implicitly using the rough hypercuboid concept. The effectiveness of the proposed algorithm, along with a comparison with other clustering algorithms, is presented on several microRNA data sets using some well-known cluster validity indices.

2 Rough Hypercuboid Approach

Let $\mathbb{U} = \{x_1, \ldots, x_j, \ldots, x_n\}$ be the finite set of n objects, and $\mathbb{C} = \{\mathcal{A}_1, \ldots, \mathcal{A}_k, \ldots, \mathcal{A}_m\}$ and $\mathbb{D}$ are, respectively, the set of condition and decision attribute in $\mathbb{U}$. $\mathbb{U}/\mathbb{D} = \{\beta_1, \ldots, \beta_i, \ldots, \beta_c\}$ represents c equivalence classes of $\mathbb{U}$ generated by the equivalence relation with the help of the decision attribute set $\mathbb{D}$. The condition attribute $\mathcal{A}_k \in \mathbb{C}$ also generates c equivalence classes $\mathbb{U}/\mathcal{A}_k = \{\delta_1, \ldots, \delta_i, \ldots, \delta_c\}$ of $\mathbb{U}$ using the equivalence relation. So c-partitions of $\mathbb{U}$ consist of sets of (cn) values $\{\mathrm{h}_{ij}(\mathcal{A}_k)\}$ that can be represented as a $(c \times n)$ matrix $\mathbb{H}(\mathcal{A}_k) = [\mathrm{h}_{ij}(\mathcal{A}_k)]$. The matrix $\mathbb{H}(\mathcal{A}_k)$ is denoted as hypercuboid equivalence partition matrix (HEPM) [11] of the condition attribute $\mathcal{A}_k$ and is expressed as

$$\mathbb{H}(\mathcal{A}_k) = \begin{pmatrix} \mathrm{h}_{11}(\mathcal{A}_k) & \mathrm{h}_{12}(\mathcal{A}_k) & \cdots & \mathrm{h}_{1n}(\mathcal{A}_k) \\ \mathrm{h}_{21}(\mathcal{A}_k) & \mathrm{h}_{22}(\mathcal{A}_k) & \cdots & \mathrm{h}_{2n}(\mathcal{A}_k) \\ \cdots & \cdots & \cdots & \cdots \\ \mathrm{h}_{c1}(\mathcal{A}_k) & \mathrm{h}_{c2}(\mathcal{A}_k) & \cdots & \mathrm{h}_{cn}(\mathcal{A}_k) \end{pmatrix}, \tag{1}$$

$$\text{where } \mathrm{h}_{ij}(\mathcal{A}_k) = \begin{cases} 1 \text{ if } \mathrm{L}_i \leq x_j(\mathcal{A}_k) \leq \mathrm{U}_i \\ 0 \text{ otherwise.} \end{cases} \tag{2}$$

Here, $\mathrm{h}_{ij}(\mathcal{A}_k) \in \{0, 1\}$ is the membership of object x_j in the ith equivalence partition β_i. Based on the concept of confusion vector [11], the HEPM can be used to identify the misclassified objects, which are bounded by the implicit hypercuboids:

$$\mathbb{V}(\mathcal{A}_k) = [\mathrm{v}_1(\mathcal{A}_k), \cdots, \mathrm{v}_j(\mathcal{A}_k), \cdots, \mathrm{v}_n(\mathcal{A}_k)] \tag{3}$$

$$\text{where } \mathrm{v}_j(\mathcal{A}_k) = \min\{1, \sum_{i=1}^{c} \mathrm{h}_{ij}(\mathcal{A}_k) - 1\}. \tag{4}$$

Let $\beta_i \subseteq \mathbb{U}$. A-lower and A-upper approximations of cluster β_i can be constructed to approximate β_i using only the information contained within $\mathcal{A}_k$:

$$\underline{A}(\beta_i) = \{x_j |\ \mathrm{h}_{ij}(\mathcal{A}_k) = 1 \text{ and } \mathrm{v}_j(\mathcal{A}_k) = 0\}; \tag{5}$$

$$\overline{A}(\beta_i) = \{x_j |\ \mathrm{h}_{ij}(\mathcal{A}_k) = 1\}; \tag{6}$$

where equivalence relation A is induced by attribute $\mathcal{A}_k$. The boundary region of β_i can be represented as

$$B_A(\beta_i) = \{\overline{A}(\beta_i) \setminus \underline{A}(\beta_i)\} = \{x_j |\ \mathrm{h}_{ij}(\mathcal{A}_k) = 1 \text{ and } \mathrm{v}_j(\mathcal{A}_k) = 1\}. \tag{7}$$

Given $< \mathbb{U}, \mathbb{C} >$ where $\mathbb{C} = \{\mathcal{A}_1, \cdots, \mathcal{A}_k, \cdots, \mathcal{A}_m\}$ is the set of condition attributes. HEPM for the set $\mathbb{C}$ is a $(c \times n)$ matrix

$$\mathbb{H}(\mathbb{C}) = \bigcap_{\mathcal{A}_k \in \mathbb{C}} \mathbb{H}(\mathcal{A}_k); \text{ where } \mathrm{h}_{ij}(\mathbb{C}) = \bigcap_{\mathcal{A}_k \in \mathbb{C}} \mathrm{h}_{ij}(\mathcal{A}_k). \tag{8}$$

This resultant HEPM $\mathbb{H}(\mathbb{C})$ can be used to construct the lower approximation and boundary region corresponding to the whole feature set $\mathbb{C}$.

3 Proposed Clustering Algorithm

The proposed RH-IT2FCM algorithm is an objective function-based clustering technique, where the sum of distances between an object and corresponding cluster prototype is minimized iteratively. Each cluster β_i is represented by a cluster center v_i, a crisp lower approximation $\underline{A}(\beta_i)$, and a IT2 fuzzy boundary region $B_A(\beta_i) = \{\overline{A}(\beta_i) \setminus \underline{A}(\beta_i)\}$, where $\overline{A}(\beta_i)$ denotes the upper approximation of cluster β_i.

3.1 Membership Function

By minimizing the objective function of traditional RFCM [6] with respect to μ_{ij}, we get

$$\mu_{ij} = \left[\sum_{k=1}^{c} \left(\frac{d_{ij}^2}{d_{kj}^2} \right)^{\frac{1}{\acute{m}_1 - 1}} \right]^{-1} \tag{9}$$

$$\text{subject to } \sum_{i=1}^{c} \mu_{ij} = 1, \forall j, \text{ and } 0 < \sum_{j=1}^{n} \mu_{ij} < n, \forall i.$$

In the RH-IT2FCM, upper and lower memberships are used to manage the uncertainties for fuzzifier $\acute{m}_1$. Footprint of Uncertainty (FOU) of fuzzifier $\acute{m}_1$ is created by two different fuzzifiers $\acute{m}_{1\_1}$ and $\acute{m}_{1\_2}$. Upper (lower) membership $\overline{\mu}_{ij}$ ($\underline{\mu}_{ij}$) is the

upper (lower) bound of the FOU that can be derived by solving the objective function with respect to $\overline{\mu}_{ij}$ and $\underline{\mu}_{ij}$ as:

$$\overline{\mu}_{ij} = \begin{cases} \left[\sum_{k=1}^{c}\left(\frac{d_{ij}^2}{d_{kj}^2}\right)^{\frac{1}{\dot{m}_{1\_1}-1}}\right]^{-1} & \text{if } \sum_{k=1}^{c}\frac{d_{ij}}{d_{kj}} < c \\ \left[\sum_{k=1}^{c}\left(\frac{d_{ij}^2}{d_{kj}^2}\right)^{\frac{1}{\dot{m}_{1\_2}-1}}\right]^{-1} & \text{otherwise.} \end{cases}$$

$$\underline{\mu}_{ij} = \begin{cases} \left[\sum_{k=1}^{c}\left(\frac{d_{ij}^2}{d_{kj}^2}\right)^{\frac{1}{\dot{m}_{1\_1}-1}}\right]^{-1} & \text{if } \sum_{k=1}^{c}\frac{d_{ij}}{d_{kj}} \geq c \\ \left[\sum_{k=1}^{c}\left(\frac{d_{ij}^2}{d_{kj}^2}\right)^{\frac{1}{\dot{m}_{1\_2}-1}}\right]^{-1} & \text{otherwise.} \end{cases} \tag{10}$$

The output of the RH-IT2FCM algorithm is an IT2 fuzzy set, so defuzzification cannot transform it to crisp set directly. Type reduction [12] is required in this case, which reduces IT2 fuzzy set to T1 fuzzy set. To get desirable clustering results with accurate cluster centers, centroid-type reducer is used. Then the type-reduced T1 fuzzy [12] values are passed through a centroid defuzzifier to obtain the crisp centers.

3.2 Cluster Prototypes

The weighted average of the objects in crisp lower approximation and probabilistic boundary is used to calculate new centroid. The effects of fuzzy membership and lower approximation and boundary region of rough sets are included in computation of a centroid. The modified centroid calculation for the RH-IT2FCM is:

$$v_i^{\text{RH-IT2FCM}} = \begin{cases} wC_1 + (1-w)D_1 & \text{if } \underline{A}(\beta_i) \neq \emptyset,\ B_A(\beta_i) \neq \emptyset \\ C_1 & \text{if } \underline{A}(\beta_i) \neq \emptyset, B_A(\beta_i) = \emptyset \\ D_1 & \text{if } \underline{A}(\beta_i) = \emptyset, B_A(\beta_i) \neq \emptyset \end{cases} \tag{11}$$

$$\text{where } C_1 = \frac{1}{|\underline{A}(\beta_i)|}\sum_{x_j \in \underline{A}(\beta_i)} x_j.$$

$\mathcal{D}_1$ can be determined by type reduction using Karnik-Mendel (KM) algorithm [12] as follows. First, all objects $x_j \in B_A(\beta_i)$ are sorted as

$$\begin{gathered} x_{11} \leqslant x_{21} \leqslant \cdots \leqslant x_{n_i 1} \\ \cdots \quad \cdots \quad \cdots \\ x_{1m} \leqslant x_{2m} \leqslant \cdots \leqslant x_{n_i m} \end{gathered} \tag{12}$$

where m is the number of features, and $n_i = |B_A(\beta_i)|$ is the number of objects in the boundary region of ith cluster. The membership μ_{ij} is initialized as $\mu_{ij} = (\overline{\mu}_{ij} + \underline{\mu}_{ij})/2$ and fuzzifier m_1 is initialized as $m_1 = (\acute{m}_{1\_1} + \acute{m}_{1\_2})/2$. Then, the ith fuzzy cluster center is calculated as

$$\acute{v}_i = \frac{\sum\limits_{x_j \in B_A(\beta_i)} (\mu_{ij})^{m_1} x_j}{\sum\limits_{x_j \in B_A(\beta_i)} (\mu_{ij})^{m_1}}. \tag{13}$$

Next step is to find the switch point s such that $x_s \leqslant \acute{v}_i \leqslant x_{s+1}$. The minimum value $\acute{v}^L$ and maximum value $\acute{v}^R$ of the ith cluster center can be found out by

$$\acute{v}_i^L = \frac{\sum\limits_{\substack{j=1 \\ x_j \in B_A(\beta_i)}}^{s} (\overline{\mu}_{ij})^{\acute{m}_1} x_j + \sum\limits_{\substack{j=s+1 \\ x_j \in B_A(\beta_i)}}^{n_i} (\underline{\mu}_{ij})^{\acute{m}_1} x_j}{\sum\limits_{\substack{j=1 \\ x_j \in B_A(\beta_i)}}^{s} (\overline{\mu}_{ij})^{\acute{m}_1} + \sum\limits_{\substack{j=s+1 \\ x_j \in B_A(\beta_i)}}^{n_i} (\underline{\mu}_{ij})^{\acute{m}_1}} \tag{14}$$

$$\acute{v}_i^R = \frac{\sum\limits_{\substack{j=1 \\ x_j \in B_A(\beta_i)}}^{s} (\underline{\mu}_{ij})^{\acute{m}_1} x_j + \sum\limits_{\substack{j=s+1 \\ x_j \in B_A(\beta_i)}}^{n_i} (\overline{\mu}_{ij})^{\acute{m}_1} x_j}{\sum\limits_{\substack{j=1 \\ x_j \in B_A(\beta_i)}}^{s} (\underline{\mu}_{ij})^{\acute{m}_1} + \sum\limits_{\substack{j=s+1 \\ x_j \in B_A(\beta_i)}}^{n_i} (\overline{\mu}_{ij})^{\acute{m}_1}} \tag{15}$$

where $\acute{m}_1 = \acute{m}_{1\_1}$, if $\left(\sum\limits_{k=1}^{c} \frac{d_{ij}}{d_{kj}} < c \text{ and } \mu_{ij} = \overline{\mu}_{ij} \right)$ *or* $\left(\sum\limits_{k=1}^{c} \frac{d_{ij}}{d_{kj}} \geq c \text{ and } \mu_{ij} = \underline{\mu}_{ij} \right)$

and $\acute{m}_1 = \acute{m}_{1\_2}$, if $\left(\sum\limits_{k=1}^{c} \frac{d_{ij}}{d_{kj}} \geq c \text{ and } \mu_{ij} = \overline{\mu}_{ij} \right)$ or $\left(\sum\limits_{k=1}^{c} \frac{d_{ij}}{d_{kj}} < c \text{ and } \mu_{ij} = \underline{\mu}_{ij} \right)$ (16)

For the computation of $\acute{v}^L$, the iteration is stopped if $\acute{v}_i^L = \acute{v}_i$ by setting the minimum value of $\acute{v}_i$ to $\acute{v}_i^L$. Otherwise setting $\acute{v}_i = \acute{v}_i^L$, the iteration is continued by searching the switch point s again. Similarly, the maximum value of $\acute{v}_i^R$ can be computed. The memberships chosen to update $\acute{v}^L$ and $\acute{v}^R$ are called μ_{ij}^L and μ_{ij}^R, respectively.

The crisp value of $\mathcal{D}_1$ can be calculated by defuzzification of $\acute{v}^L$ and $\acute{v}^R$ as $\mathcal{D}_1 = (\acute{v}^L + \acute{v}^R)/2$. The new cluster centroids are computed as per (11). The above steps are repeated until the cluster centers are almost same as that of the previous iteration. The resulting cluster centroids are dependent on the parameters w and $(1 - w)$, and fuzzifiers $\acute{m}_{1\_1}$ and $\acute{m}_{1\_2}$ manage their relative influence.

3.3 Algorithm

The main steps of the RH-IT2FCM are as follows:

1. Assign iteration counter $t = 1$. Assign initial cluster prototypes $\acute{v}_i^{(t)}$, $i = 1, 2, \ldots, c$. Choose the values for fuzzifiers $\acute{m}_{1\_1}$ and $\acute{m}_{1\_2}$. Set a very small nonzero positive value to ε.
2. Put each object x_j in the group that has the closest centroid.
3. Construct the HEPM $\mathbb{H}(\mathcal{A}_l)$ for each feature $\mathcal{A}_l \in \mathbb{C}$ according to (1), considering the cluster labels of x_j, $\forall j = 1, 2, \ldots, n$.
4. Construct resultant HEPM $\mathbb{H}(\mathbb{C})$, considering all features $\mathcal{A}_l \in \mathbb{C}$.
5. Calculate the lower approximation $\underline{A}(\beta_i)$ and boundary region $B_A(\beta_i)$ using (5) and (7).
6. Compute $\overline{\mu}_{ij}$ and $\underline{\mu}_{ij}$ for the objects located in boundary regions for c clusters using (10).
7. Objects are removed from the boundary region $B_A(\beta_i)$, if $\mu_j < \widetilde{\mu}_j$, where $\mu_j = \max_i\{\mu_{ij}\}$, and $\mu_{ij} = (\overline{\mu}_{ij} + \underline{\mu}_{ij})/2$. $\widetilde{\mu}_j$ is the average of the memberships μ_j of all of the objects residing in boundary region $B_A(\beta_i)$.
8. Compute new centroid $\acute{v}_i^{(t+1)}$ as per (11), where the centroid for the objects in boundary region is calculated using Karnik-Mendel (KM) algorithm [12].
9. Repeat steps 2 to 8, by increasing t, until no more change is observed, that is, $\|\acute{v}^{(t+1)} - \acute{v}^{(t)}\| < \varepsilon$.

Type reduction from IT2 to T1 can be done as follows: $\mu_{ij} = (\mu_{ij}^L + \mu_{ij}^R)/2$, where μ_{ij}^L and μ_{ij}^R are the left and right memberships calculated using KM algorithm. Finally, objects are assigned to the clusters using the maximum membership, i.e., $x_j \in v_i$ if $\mu_{ij} = \max_k\{\mu_{kj}\}$.

4 Experimental Results

The results obtained by the proposed RH-IT2FCM algorithm are compared extensively with that of different c-means algorithms on three miRNA data sets. The algorithms compared are hard c-means (HCM) [1], fuzzy c-means (FCM) [13], rough c-means (RCM), rough-fuzzy c-means (RFCM) [6], robust rough-fuzzy c-means (rRFCM) [14], and rough hypercuboid-based FCM (RHFCM). The results are analyzed with respect to Silhouette index, DB index, β index, Xie-Beni index, and execution time. All the algorithms are programmed in C language and executed in Ubuntu platform. The input parameters are kept fixed for all executions, and values of fuzzifier $\acute{m}_1 = 2.0$. $\acute{m}_{1\_1}$ and $\acute{m}_{1\_2}$ are changed taking the values from the set {1.1 1.5 2.0 2.5 3.0 3.5 4.0 4.5}, for all values of $\acute{m}_{1\_2} \geq \acute{m}_{1\_1}$. For rough-fuzzy clustering, the weight parameter $w = 0.99$ is used.

4.1 Performance of Various C-Means Algorithms

To establish the superiority of the proposed RH-IT2FCM over the rough-fuzzy c-means (RFCM) [6], robust rough-fuzzy c-means (rRFCM) [14], rough hypercuboid-based fuzzy c-means (RHFCM), and other c-means algorithms, extensive experimentation is carried out on three miRNA data sets. The experimental results are represented in Table 1. The bold values in Table 1 signify the best values. Out of 12 cases, the RH-IT2FCM algorithm outperforms the other methods in 11 cases.

The results presented in Table 1 lead to the following conclusions.

1. The proposed clustering algorithm is superior to other fuzzy and rough-based c-means clustering algorithms with respect to all cluster validity indices used.
2. The proposed RH-IT2FCM algorithm produces better results with lesser time than that generated using existing robust rough-fuzzy c-means, for all the data sets and quantitative indices. The hypercuboid approach helps to partition the objects in lower approximation and boundary region without any user-specified parameter, giving better results in terms of clustering.
3. The RH-IT2FCM algorithm also produces better results than that obtained using rough hypercuboid-based fuzzy c-means, for all the data sets and quantitative indices. Inclusion of IT2 fuzzy approach helps to manage the uncertainty present in the fuzzy membership function.

The best performance of the RH-IT2FCM, in terms of Silhouette, DB, β, and Xie-Beni indices, is achieved as the rough hypercuboid approach used in the RH-IT2FCM computes the lower approximation and boundary region implicitly for the clusters without any threshold value; and the concept of crisp lower approximation and interval type-2 fuzzy boundary of the RH-IT2FCM algorithm deals with uncertainty, vagueness, and incompleteness in cluster definition, while the IT2 fuzzy approach helps to manage the uncertainty present in the fuzzy membership function.

Table 1 Performance of various c-means algorithms on three microRNA data sets

MicroRNA data sets	Different algorithms	Different cluster-validity indices				Execution time
		Silhouette index	DB index	β index	Xie-Beni index	
GSE50505	HCM	0.411012	0.715626	3.097629	0.122234	205
	FCM	0.365385	0.842116	3.051504	0.184100	752
	RCM	0.444451	0.856552	2.729572	0.165834	50
	RFCM	0.397289	0.764749	2.783327	0.136991	176
	rRFCM	0.516692	**0.313635**	8.063505	0.051250	174
	RHFCM	0.411012	0.689686	3.091021	0.115353	31
	RH-IT2FCM	**0.576592**	0.383510	**8.869973**	**0.009269**	90
GSE52677	HCM	0.046198	4.366908	1.239672	2.245455	168
	FCM	0.046198	4.366908	1.239672	2.245455	208
	RCM	0.121243	3.088299	1.132286	1.655467	290
	RFCM	0.066989	2.845340	1.126628	1.517413	126
	rRFCM	0.093178	2.103604	1.703359	1.102970	456
	RHFCM	0.268940	1.550170	1.718763	0.734185	33
	RH-IT2FCM	**0.388768**	**0.853854**	**4.395165**	**0.089416**	131
GSE71107	HCM	0.377091	0.762593	9.018855	0.145465	651
	FCM	0.343229	1.247167	8.559175	0.376418	2454
	RCM	0.447050	0.876967	8.451245	0.351707	1258
	RFCM	0.359525	0.836476	8.362779	0.287403	864
	rRFCM	0.434765	0.489441	22.11733	0.134818	1475
	RHFCM	0.471422	0.419962	6.170177	0.077129	217
	RH-IT2FCM	**0.734994**	**0.168680**	**45.40408**	**0.009920**	1245

4.2 Functional Consistency of Clustering Results

To predict microRNA target genes for the microRNA clusters produced by different clustering algorithms, a microRNA target prediction algorithm is used, called DIANA microT. The genes which are targeted by minimum t percentage of microRNAs in each microRNA cluster are analyzed further, where t is varied from 50 to 100. Figure 1 depicts the comparative results of the HCM, FCM, RCM, RFCM, rRFCM, RHFCM, and the proposed RH-IT2FCM algorithms, on the basis of cluster frequency or final annotation ratio (FAR), for the Molecular Function (MF), Biological Process (BP), and Cellular Component (CC) ontologies for three microRNA expression data sets mentioned above. All the results presented in the figure confirm that the proposed RH-IT2FCM mostly provides higher or comparable FARs than that produced by the other c-means algorithms. From the results studied, it is observed that, out of 22 comparisons each, the RH-IT2FCM attains a highest FAR than that generated using HCM, FCM, and RCM algorithms in 20, 19, and 20 cases,

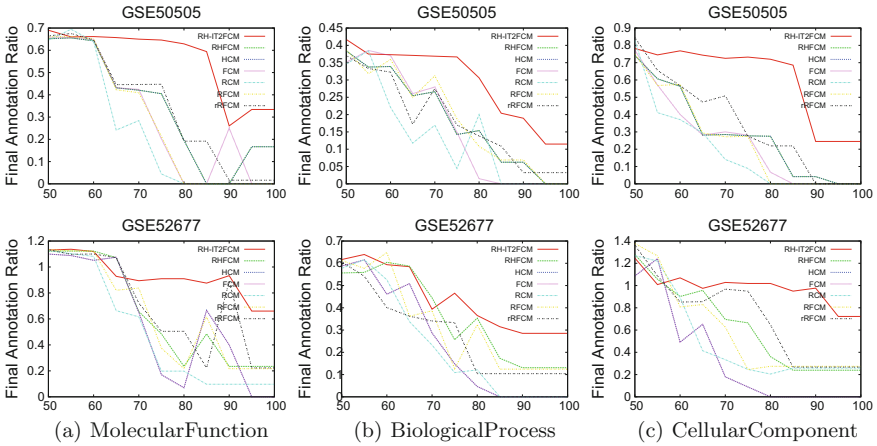


Fig. 1 Biological annotation ratios produced by different algorithms

respectively, for the MF, BP, and CC ontologies. The results also confirm that the RH-IT2FCM method mostly provides higher or comparable FAR than that obtained using the RFCM, rRFCM, and RHFCM algorithms. Out of 22 cases each, the RH-IT2FCM method provides a higher FAR in 21, 21, and 21 cases for the MF, BP, and CC ontologies, respectively. Overall, the RH-IT2FCM method provides a higher FAR in 19, 18, and 19 cases for the MF, BP, and CC ontologies, respectively, out of 22 cases each. Hence, the miRNA clusters generated by RH-IT2FCM algorithm are functionally more compact than those obtained using other algorithms, as the other algorithms include noise or irrelevant miRNAs in the clusters.

5 Conclusion

The contribution of this paper is twofold, namely,

1. the development of a new algorithm for clustering miRNAs, integrating judiciously the c-means algorithm, rough hypercuboid approach, probabilistic memberships of fuzzy sets, and IT2 fuzzy approach, to overcome the problems of the uncertainty present in miRNA data, and the uncertainty in determining the fuzzy membership function;
2. demonstrating the effectiveness of the proposed algorithm, along with a comparison of other related algorithms, on three miRNA data sets using some standard cluster validity indices, and the functional consistency of the resultant miRNA clusters.

The algorithm is formulated by maximizing the utility of both rough and fuzzy sets for knowledge discovery tasks. The proposed clustering method is found to pro-

vide the best performance in 91.7% cases than the other clustering methods reported. Moreover, the performance of proposed algorithm is significantly better than other algorithms, irrespective of the miRNA data sets and quantitative measures used, and generates relevant and functionally consistent miRNA clusters in lesser or comparable time.

References

1. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs, NJ (1988)
2. Domany, E.: Cluster Analysis of gene expression data. J. Stat. Phys. **110**(3–6), 1117–1139 (2003)
3. Pawlak, Z.: Rough Sets: Theoretical Aspects of Reasoning About Data. Kluwer, Dordrecht, The Netherlands (1991)
4. Maji, P., Garai, P.: Fuzzy-Rough simultaneous attribute selection and feature extraction algorithm. IEEE Trans. Syst. Man Cybern. Part B Cybern. **43**(4), 1166–1177 (2013)
5. Maji, P., Garai, P.: Simultaneous feature selection and extraction using fuzzy rough sets. In: Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), pp. 115–123, Dec 2012
6. Maji, P., Pal, S.K.: Rough set based generalized fuzzy C-means algorithm and quantitative indices. IEEE Trans. Syst. Man Cybern. Part B Cybern. **37**(6), 1529–1540 (2007)
7. Zadeh, L.A.: The concept of a linguistic variable and its application to approximate reasoning-1. Inf. Sci. **8**, 199–249 (1975)
8. Rhee, F., Hwang, C.: A type-2 fuzzy C-means clustering algorithm. In: Joint 9th IFSA World Congress and 20th NAFIPS International Conference, vol. 4, pp. 1926–1929, July 2001
9. Hwang, C., Rhee, C.: Uncertain fuzzy clustering: interval type-2 fuzzy approach to C-means. IEEE Trans. Fuzzy Syst. **15**(1), 107–120 (2007)
10. Maji, P., Garai, P.: IT2 fuzzy-rough sets and max relevance-max significance criterion for attribute selection. IEEE Trans. Cybern. **45**(8), 1657–1668 (2015)
11. Maji, P.: Rough hypercuboid approach for feature selection in approximation spaces. IEEE Trans. Knowl. Data Eng. **26**(1), 16–29 (2014)
12. Mendel, J.M., Karnik, N.N.: Centroid of a type-2 fuzzy set. Inf. Sci. **132**(1), 195–220 (2001)
13. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithm. Plenum Press, New York (1981)
14. Maji, P., Paul, S.: Robust rough-fuzzy C-means algorithm: design and applications in coding and non-coding RNA expression data clustering. Fundam. Inf. **124**, 153–174 (2013)

A Novel Algorithm for Network Anomaly Detection Using Adaptive Machine Learning

D. Ashok Kumar and S. R. Venugopalan

Abstract Threats on the Internet are posting high risk to information security and network anomaly detection has become an important issue/area in information security. Data mining algorithms are used to find patterns and characteristic rules in huge data and this is very much used in Network Anomaly Detection System (NADS). Network traffic has several attributes of qualitative and quantitative nature, which needs to be treated/normalized differently. In general, a model is built with the existing data and the system is trained with the model and then used to detect intrusions. The major and important issue with such NADS is that the network traffic changes over time; in such cases, the system should get trained automatically or retrained. This paper presents an adaptive algorithm that gets trained according to the network traffic. The presented algorithm is tested with Kyoto University's 2006+ Benchmark dataset. It can be observed that the results of the proposed algorithm outperform all the known/commonly used classifiers and are very much suitable for network anomaly detection.

Keywords Intrusion · Anomaly · Network traffic · Normalization Performance metrics · Adaptive algorithm · Kyoto 2006+ · Naïve Bayes classification

1 Introduction

Internet has brought huge potential for business and on the other hand, it poses lots of risk for the business. Internet is a global public network [1]. Intrusion is a deliberate, unauthorized, illegal attempt to access, manipulate, or take possession of

D. Ashok Kumar (✉)
Department of Computer Science, Govt. Arts College, Tiruchirappalli, Tamilnadu, India
e-mail: akudaiyar@yahoo.com

S. R. Venugopalan
Aeronautical Development Agency (Ministry of Defence, GoI), Bengaluru 560017, India
e-mail: venu_srv@yahoo.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_7

information system to render them unreliable or unusable. Intrusion detection is the process of identifying various events occurring in a system/network and analyzing them for the possible presence of intrusion. Intrusion Detection Systems (IDS) can be classified into three types based on the method on which intrusion is detected namely signature-based, anomaly-based, and hybrid. Statistical methods and clustering are used for anomaly detection systems [1]. The availability of higher bandwidth and sophisticated hardware and software, the need to detect intrusions in real-time, and the adaptation of the detection algorithm to the ever-changing traffic pattern are a big challenge. IDS should adapt to the traffic behaviors and learn automatically. In this paper, an algorithm is proposed for network anomaly detection. The results, i.e., performance metrics of the experiment, are encouraging. The proposed algorithm can detect new/unknown attacks and can learn and adapt automatically based on the network traffic.

The organization of the paper is as follows: Sect. 2 gives the background and the literature surrounding IDS with necessary performance metrics. The problem description and the algorithm development are discussed in Sect. 3. In Sect. 4, the dataset, data preprocessing, data normalization, and the training and test dataset generation used in this study are discussed. The experiment and the results are discussed in Sect. 5. Conclusions and future work are in given in Sect. 6.

2 Background and Related Work

Panda M. et al. proposed Naïve Bayes for Network Intrusion Detection and found that the performance of Naïve Bayes is better in terms of false-positive rate, cost, and computational time for KDD '99 datasets, and same was compared with backpropagation neural networks approach [2]. Jain et al. in their work have combined information gain with Naïve Bayes for improving the attack detection and have observed higher detection rate and reduced false alarm [3]. Muda Z. et al. in their work have used k-means to cluster the data and used Naïve Bayes classifier to classify the KDD Cup99 [4] data and have achieved better performance than Naïve Bayes classifier [5]. They have achieved 99.7% accuracy, a detection rate of 99.8%, and 0.5 false alarm rate.

FVBRM model is proposed by the authors of [6] for feature selection and compared it with other selection methods by reducing the features of the dataset and then classifying with Naive Bayes classifier. There is no mention about how the qualitative and quantitative attributes are treated. The authors of [7] have compared the results of Naïve Bayes algorithm with decision tree and concluded that from the performance point of view Naïve Bayes provides competitive results for KDD 99 [8] dataset. K-means clustering algorithm was applied for intrusion detection and concluded that k-means method is very efficient in partitioning huge dataset and has better global search ability [9, 10]. K-means clustering is a good unsupervised algorithm used to find out structured patterns in the data but the computational complexity is high for its application in intrusion detection. A novel density based

k-means cluster was proposed for signature-based intrusion detection [11] where results show improved accuracy and detection rate with reduced false-positive rate. It is not very clear that which normalization technique was used and how the discrete and continuous data was treated. Sharma et al. [12] proposed k-means clustering via Naïve Bayes for KDD Cup '99 dataset. This approach outperforms the Naïve Bayes in terms of detection rate and higher false positives which is a concern.

S.M. Hussein et al. in their work compared the performance of Naïve Bayes, Bayes Net, and J48graft, and recorded that Naïve Bayes performs better in terms of rate of detection and time to build model, whereas J48 was better in terms of false alarm rate [13]. Earlier works which were reviewed in this section tried in achieving higher performance with the help of preprocessing/feature reduction and have achieved performance improvements. The study of the existing literature reveals the need for a novel algorithm to detect unknown attacks because they have not considered the following points: (a) Ever-changing network traffic/speed, new attacks, and the need for the algorithm to adapt itself and learn/get trained automatically from the changing traffic; (b) The ability of the algorithms/methods described in the literature to perform well for datasets other than the tested ones. The algorithms were tested with the only one dataset; (c) Either attack or normal data is used for training and not both; (d) Network traffic data contains features that are qualitative or quantitative nature and has to be treated differently and have to use different preprocessing/normalization technique; and (e) Earlier works have measured accuracy, detection rate, and false alarm rate only as a performance measure which may not be sufficient; measures such as F-score and sensitivity are required for evaluating an algorithm/method.

2.1 Metrics for Intrusion Detection Performance

The choice of NADS for a particular environment is a general problem, represented precisely as intrusion detection system's evaluation [14]. For an anomaly detection system, False Alarm Rate (FAR) and the Detection Rate (DR) are basic factors and their trade-off can be analyzed with Receiver Operating Characteristic (ROC) curve. The above-mentioned basic factors FAR and DR are not sufficient to evaluate the performance of IDS [15]. So the evaluation of IDS should take into account the environment where the IDS is being deployed, its maintenance costs, operating environments, likelihood of attacks, cost toward false alarm and missed detections, etc. [14]. The following section explains the performance metrics, which needs to be considered while deploying/deciding on IDS/anomaly detection system and these measures are used for evaluation of the algorithm proposed. Attacks that are detected correctly as attacks are referred as True Positives (TP) and normal connections detected as normal connections are True Negatives (TN). The following Table 1 is the general confusion matrix used in intrusion detection evaluation. The values in the matrix represent the performance of the prediction algorithm. TP rate determines the security requirement and the number of FP's determines the

Table 1 Confusion matrix

Confusion matrix		Predicted value	
		Attack	Normal
Actual value	Attack	True Positives (TP)	False Negatives (FN)
	Normal	False Positives (FP)	True Negatives (TN)

Table 2 Performance measures used to evaluate IDS

S. no.	Performance metric	Description	Formula
1.	Detection Rate/Positive Prediction Value/Precision	Proportion of the predicted positives which are actual positive (or) fraction of test data detected as attack which is actually an attack	(TP + FP)
2.	Accuracy	Measure to test the overall accuracy. It can be delineated as the percentage of correct prediction among the whole dataset	(TP + FP + FN + TN)
3.	False Alarm Rate	False-positive rate (FPR) also known as false alarm rate (FAR) refers to the proportion of normal packets being falsely detected as malicious	(FP + TN)
4.	Sensitivity/True Positive Rate/Recall	The fraction of attack class which is correctly detected (or) proportion of actual positives which are predicted as positives	(TP + FN)

usability of the IDS. There is always a trade-off between the two metrics, precision and recall. For an IDS to be effective, the FP and FN rates should be minimized and accuracy, and TP and TN rates to be maximized [16].

Table 2 gives the details about the various performance measures for the evaluation of IDS.

F-Score

The harmonic mean between precision and recall is called as F-score/F-measure. F-score is considered as a measure of the accuracy of a test. Good IDS performance is achieved by improving both precision and recall. Both precision and recall are considered for computing F-score. An F-score of 1 is considered as best and 0 as worst:

$$F - \text{Score} = \frac{2 * \text{P} * \text{R}}{\text{P} + \text{R}} \tag{1}$$

3 Problem Description and Approach/Algorithm Development

Supervised algorithms significantly outperform unsupervised algorithms in detecting known attacks. For those problems where the test data is drawn from different distributions, semi-supervised learning methods offer a promising future [17]. The dramatic increase in the speed of the networks has made the existing policies and network anomaly intrusions detection systems vulnerable to intrusion than ever before. Thus, making the existing IDS useless unless they adapt to the new trends, i.e., adapt to the ever-changing network traffic and learn automatically. Adaptive Network Anomaly Detection Algorithm (ANADA) proposed in this study uses labeled dataset for initial learning and adapts itself to the changing traffic patterns. The proposed ANADA algorithm used simple statistical measures such as mean, median, and norm (distance measure). This algorithm uses normalized data, i.e., the normalization of training data is described in the data preprocessing section. The uniqueness of the algorithm is given below:

- The algorithm uses both attack and normal data for training;
- The algorithm adapts itself the new traffic by modifying the training dataset with the test dataset;
- At each test instance, the algorithm decides whether the test data is worth being included/replaced with an instance of training data;
- The algorithm is very simple and can be easily parallelized for performance improvements; and
- This algorithm uses a new distance measure, i.e., 0.8 norm (given in Eq. 2).

3.1 Adaptive Network Anomaly Detection Algorithm (ANADA)

Input: Training dataset and testing dataset: a—attack training dataset; n—normal training dataset; and t—testing dataset.

Output: Anomaly detection performance metrics such as detection rate, FAR, sensitivity, F-score, etc.

Generate initial population/training dataset that has equal number (5000) of attacks and normal traffic features.

Training Phase: The training dataset is grouped based on the label as attack and normal sessions. 5000 attack records and 5000 normal records are used for training. Find the centroid of the attack class and normal class. For numerical attributes, the mean (or) average is calculated and for the categorical attributes, median is calculated. The centroid will be a set of values.

Testing Phase: For each record in the testing data, the following steps are followed:

BEGIN

1. *Initialize the necessary variables such as counters and loop index etc.*
2. *Read the attack and normal traffic data. // attack data is referred as a[5000] [7] and normal data as n[5000] [7].*
3. *Evaluate mean for first 12 attributes and median for next 2 attributes for both attack and normal data //ma referred as mean of train attack data and mn referred as mean of train normal data.*
4. *Read the test data // test data is referred as t[5000] [7] 15th column is the actual label and 16th column will be used for computed label.*
5. *Compute the distance between the test data and the centroid of the attack/normal dataset using 0.8-norm as given in Eq. 2:*

$$|X| = \sqrt[0.8]{\sum_{k=1}^{n} |ai - ti|^{0.8}} \tag{2}$$

6. *If the test data is closer to normal centroid and the distance between test data and normal centroid is less than 1.5 times of the distance between the normal and attack centroid, then it is labeled as normal else an attack.*
7. *After labeling the test data, decision has to be made whether to replace the test data with the training data.*
8. *If the new test data is attack/normal, the decision has to be made whether the new data has to be replaced with the attack/normal training data or not. This is done by calculating the distance between the test data and the attack/normal centroid and the ith (counter used for replacement) row of attack data and the centroid of the attack/normal. The distance is calculated using 0.8-norm as given in Eq. 2. If the new test data is closer to the centroid than the ith data, then replace the ith data with the new one.*
9. *Repeat the above steps for all the test data. The algorithm is given in the next Sect. 3.1.*
10. *Calculate the TP, TN, FP, FN, sensitivity, specificity, FAR, accuracy, detection rate, F-score, etc.*

END *//end of algorithm.*

4 Datasets for Experimentation

In this paper, the publicly available dataset Kyoto 2006+ datasets is used for experimentation.

4.1 Kyoto 2006+ Dataset

Kyoto 2006+ [18] dataset is a network intrusion evaluation/detection dataset which was collected from various honeypots from November 2006 to August 2009. Real network traffic traces were captured in this dataset. This data has 24 statistical features, which include 14 features which were there in KDDCUP '99 dataset and additional ten features for effective investigation. This study uses August 31, 2009, data and has used the first 14 features (conventional features) and the label which indicates whether the record is an attack or normal. As the study does not distinguish between the known and unknown attack, both are represented as attack only. The unknown attacks in this dataset are very minimal and that is also another reason for not distinguishing known and unknown attack.

4.2 Data Preprocessing

Raw data needs to be preprocessed before fed into any learning model and the most used technique is normalization [19]. Network traffic data contains features that are qualitative or quantitative nature and have to be treated differently. The values of attributes with high values can dominate the results than the attributes with lower values [20]. The dominance can be reduced by the process of normalization, i.e., scaling the values within certain range. The quantitative attributes can be normalized by various techniques, such as (1) mean-range normalization, (2) frequency normalization, (3) maximize normalization, (4) rational normalization, (5) ordinal normalization, (6) softmax scaling [21], and (7) statistical normalization, whereas applying the above normalization techniques for qualitative data will not be meaningful. For qualitative data, the general approach is to replace the values with numerical values. Though this seems simpler, it does not consider the semantics of the qualitative attributes. In this study, the following probability function is used for normalizing the qualitative data [2, 20]:

$$\mathrm{fx}\ (\mathrm{x}) = \Pr\ (\mathrm{X} = \mathrm{x}) = \Pr(\{\mathrm{s} \in \mathrm{S}: \mathrm{X}(\mathrm{s}) = \mathrm{x}\}) \tag{3}$$

Based on the above equation, the qualitative data are converted into quantitative data in the range of [0–1]. In this study, for quantitative attributes, mean-range normalization is used [22]:

$$Xi = \frac{(vi) - \min(vi)}{\max(vi) - \min(vi)} \tag{4}$$

The reason for choosing the mean range (for quantitative attributes) and probability function (for qualitative attributes) is because this normalization technique yields better results in terms of time and classification rate [2 and 8]. There are two

qualitative attributes, i.e., flag and service; and all the other 12 attributes are quantitative. The mean-range normalization is applied for quantitative attributes and the above probability function is used for qualitative attributes.

4.3 Dataset Generation for Training and Testing

The framework used in the study uses both normal and malicious (attack) data for training. In general, the system is trained using either normal data or attack data. This is one of the unique characteristics of the algorithm which makes it suitable for adaptive learning, i.e., the system is automatically trained based on the testing/network traffic data. The data pertaining to date August 31, 2009 of Kyoto 2006+ dataset is used for this study and this dataset has 134665 records, out of which 44257 (32.9%) are normal and the 90408 (67.1%) are attack data records. There were a lot of duplicate records (42.2%) which were removed before the experimentation. *From the above statistical information, it can be observed that the attack data dominates the dataset which is not a general case and there are a lot of duplicates.*

In this study, the procedure was devised in selecting the testing/training data in such a way that the above observations do not dominate the detection procedure and this can be used for all the datasets. In this study, the training dataset consists of 5000 attack and 5000 normal records. Four sets of testing records were generated in the following manner for Kyoto 2006+ dataset. These records were chosen in random using SPSS Statistics V20 after removing the duplicates.

Dataset1 (Test Case-1) consists of 10000 records of which 10% are attack and the rest 90% are normal records.

Dataset2 (Test Case-2) consists of 10000 records of which 20% are attack and the rest 80% are normal records.

Dataset3 (Test Case-3) consists of 20000 records of which 10% are attack and the rest 90% are normal records.

Dataset3 (Test Case-4) consists of 20000 records of which 20% are attack and the rest 80% are normal records.

The reason for choosing the above configuration was that in general, the number of attacks will not be more than 20% of the records.

5 Experimental Results and Discussions

The ANADA described earlier in this study is implemented using Matlab version 7.12.0.635 (R2011a). The experiments were carried out on a system with Intel Core i3 2.53 Ghz CPU and 4 GB memory running Window 8 Professional 64-bit Operating System. Microsoft Office Professional Plus 2010 & SPSS Statistics V20 were used for data preprocessing.

Table 3 IDS performance comparison of ANADA with Naïve Bayes (Kyoto 2006+)

Kyoto dataset		Detection rate	Accuracy	False alarm rate	F-score
Test Case-1	NB Train and Test	0.7229	0.9616	0.0426	0.8388
	NB 10-Fold	0.7223	0.9615	0.0423	0.8380
	ANADA	0.8861	0.9773	0.0127	0.8866
Test Case-2	NB Train and Test	0.8499	0.9646	0.0441	0.9187
	NB 10-Fold	0.8512	0.9642	0.0435	0.9175
	ANADA	0.9402	0.9750	0.0149	0.9373
Test Case-3	NB Train and Test	0.7244	0.9619	0.0422	0.8398
	NB 10-Fold	0.7266	0.9621	0.0417	0.8404
	ANADA	0.8085	0.9727	0.0251	0.8744
Test Case-4	NB Train and Test	0.8484	0.9641	0.0446	0.9176
	NB 10-Fold	0.8525	0.9644	0.0430	0.9178
	ANADA	0.9336	0.9666	0.0159	0.9148

Kyoto 2006 dataset is preprocessed as given above and the training data was fed to the algorithm for learning. There are four test cases namely test-case1, test-case2, etc. The test cases are fed one by one and the results are recorded. The results are given in Table 3 and Fig. 1. Table 3 clearly depicts the various anomaly detection evaluation performance measures of ANADA algorithm for Kyoto 2006+ dataset. The results need to be compared with the other techniques. Naïve Bayes classification was used because of the reason that it is a simple classification scheme and

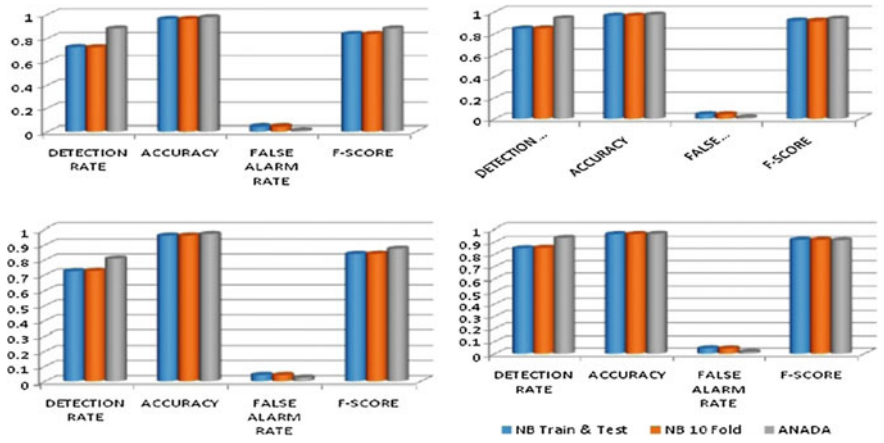


Fig. 1 Performance comparison of ANADA with NB and NB 10 Fold (Kyoto 2006+ dataset)

provides better results in terms of detection rate and FAR. Naïve Bayes is a supervised algorithm based on Bayes' theorem with the "Naïve" assumption that the features are strongly independent and mathematically this is given in Eq. 5.

$$P(X1, \ldots, Xn|Y) = \pi P(Xi|Y) \tag{5}$$

Naïve Bayes model was built using the same training set with 5000 attack and 5000 normal vectors. All the four test cases were re-evaluated with the model built and the results are tabulated. In addition to above, the test cases were evaluated using Naïve Bayes (NB) 10-fold cross-validation. The cross-validation is a process of repeatedly carrying out the experiment 10 times so that each subset is used as test set at least once. This is used to estimate the accuracy and this has been found to be effective when there is sufficient data. The results of the NB Train and Test, NB 10-fold cross-validation, and ANADA are given in Table 3 and the same is depicted as graphs in Fig. 1.

From the above table, it can be clearly observed that DR and accuracy of ANADA are higher in all the cases and F-score of ANADA is also higher in all the cases except for test case-4 which is marginally low. False Alarm Rate (FAR) is lower than NB's Train and Test and 10-fold cross-validation in all the cases which qualifies the usability of the algorithm.

6 Conclusions and Future Work

In this study, a novel adaptive algorithm has been proposed. The proposed method uses the labeled dataset for training but can adapt/learn itself and can detect new attacks. The performance measures of the algorithm can still be improved by combining this algorithm with feature weights. The algorithm has good potential to be parallelized. The future work shall focus on parallelizing the algorithm using GPGPU processors for achieving performance as energy efficiency has become the prime concern for the computer industry. Different sensors for different protocol types can be used for performance improvements. The authors are working on improving the algorithm and modifying it for flow-based anomaly detection.

References

1. https://www.sans.org/reading-room/whitepapers/detection/intruion-detection-systems-definition-challenges-343. Accessed on 06 Jan 2016
2. Panda, M., Patra, M.R.: Network intrusion detection using naive bayes. Int. J. Comput. Sci. Netw. Secur. **7**(12), 258–263 (2007)
3. Jain, M., Richariya, V.: An improved techniques based on Naïve Bayesian for attack detection. Int. J. Emerg. Technol. Adv. Eng. **2**(1), 324–331 (2012)

4. The UCI KDD Archive: KDD Cup 1999 Data, Information and Computer Science, University of California, Irvine. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html (1999). Accessed 2 February 2014
5. Muda, Z., Yassin, W., Sulaiman, M.N., Udzir, N.I.: A K-Means and Naive Bayes learning approach for better intrusion detection. Inf. Technol. J. **10**(3), 648–655 (2011)
6. Mukherjee, S., Sharma, N.: Intrusion detection using naive Bayes classifier with feature reduction. Procedia Technol. **4**, 119–128 (2012)
7. Amor, N.B., Benferhat, S., Elouedi, Z.: Naive bayes vs decision trees in intrusion detection systems. In: Proceedings of the 2004 ACM Symposium on Applied Computing, pp. 420–424 (2004)
8. MIT Lincoln Lab., Information Systems Technology Group: The 1998 Intrusion detection off-Line Evaluation Plan. http://www.ll.mit.edu/ideval/files/id98-eval-ll.txt (1998)
9. Münz, G., Li, S., Carle, G.: Traffic, Anomaly detection using K-Means Clustering. In: GI/ITG Workshop MMBnet, Sept 2007
10. Jianliang, M., Haikun, S., Ling, B.: The application on intrusion detection based on k-means cluster algorithm. In: International Forum on Information Technology and Applications, 2009. IFITA'09, pp. 150–152 (2009)
11. Randeep, B., Sharma, N.: A novel density based K-Means clustering algorithm for intrusion detection. In: J. Netw. Commun. Emerg. Technol. **3**(3), 17–22 (2015)
12. Sharma, S.K., Pandey, P., Tiwari, S.K., Sisodia, M.S.: An improved network intrusion detection technique based on K-means clustering via Naïve Bayes classification. In: 2012 International Conference on Advances in Engineering, Science and Management (ICAESM), proceedings, 30–31 Mar 2012. IEEE, Piscataway, NJ (2012)
13. Hussein, S.M., Ali, F.H.M., Kasiran, Z.: Evaluation effectiveness of hybrid IDs using snort with naive Bayes to detect attacks. In: 2012 Second International Conference on Digital Information and Communication Technology and it's Applications (DICTAP). IEEE (2012)
14. Thomas, C: Performance Enhancement of Intrusion Detection Systems using Advances in Sensor Fusion, Phd Thesis. Supercomputer Education and Research Center, Indian Institute of Science Bangalore, India (2009)
15. Gaffney Jr., J.E., Ulvila, J.W.: Evaluation of intrusion detectors: a decision theory approach. In: 2001 IEEE Symposium on Security and Privacy, 2001. S&P 2001. Proceedings, pp. 50–61. IEEE (2001)
16. Mokarian, A., Faraahi, A., Delavar, A.G.: False positives reduction techniques in intrusion detection systems-a review. Int. J. Comput. Sci. Netw. Secur. (IJCSNS) **13**(10), 128 (2013)
17. Laskov, P., Düssel, P., Schäfer, C., Rieck, K.: Learning intrusion detection: supervised or unsupervised? In: Image Analysis and Processing–ICIAP 2005, 1 Jan 2005, pp. 50–57. Springer, Berlin (2005)
18. Song, J., Takakura, H., Okabe, Y., Eto, M., Inoue, D., Nakao, K.: Statistical analysis of honeypot data and building of Kyoto 2006+ dataset for NIDS evaluation. In: Proceedings of the 1st Workshop on Building Analysis Datasets and Gathering Experience Returns for Security, Salzburg, 10–13 Apr 2011, pp. 29–36. ACM 2011 (2011). http://dx.doi.org/10.1145/1978672.1978676
19. Ammar, A.: Comparison of feature reduction techniques for binominal classification of network traffic. J. Data Anal. Inf. Process. (2015) http://dx.doi.org/10.4236/jdaip.2015.32002
20. Ihsan, Z., Idris, M.Y., Abdullah, A.H.: Attribute normalization techniques and performance of intrusion classifiers: a comparative analysis. Life Sci. J. **10**(4), 2568–2576 (2013)
21. Chavez, A.R., Hamlet, J., Lee, E., Martin, M., Stout, W.: Network Randomization and Dynamic Defence for Critical Infrastructure Systems, Sandia National Laboratories, New Mexico. SAN2015-3324 (2015)
22. Wang, W., Zhang, X., Gombault, S., Knapskog, S.J.: Attribute normalization in network intrusion detection. In: 2009 10th International Symposium on Pervasive Systems, Algorithms, and Networks (ISPAN), 14 Dec 2009, pp. 448–453. IEEE (2009)

Recognition of Odia Conjunct Characters Using a Hybrid ANN-DE Classification Technique

Mamata Nayak and Ajit Kumar Nayak

Abstract A good amount of research has been done on optical character recognition (OCR) for different languages and many research articles have been published on this topic throughout last few decades. However, not much works have been reported for conjunct character recognition of Indian languages although there are 12 major scripts used by the peoples of India. Odia is an Indian language and is used by a majority of population in Odisha. The work presented here focuses on recognition of Odia conjunct characters, which uses an optimization technique embedded within the soft computing paradigm, i.e., a neural network. The experimental results show significant improvement on recognition rate of conjunct characters compared to traditional neural network approach.

Keywords Artificial neural network (ANN) · Optical character recognition (OCR) Differential evolution (DE) · Odia conjuncts

1 Introduction

Optical character recognition deals in recognition and classification of characters from an image. Many works have been published toward recognition of Arabic, Chinese, Roman, and Japanese scripts and also for few Indian languages such as Hindi [1], Bangla, Devanagari, Gujarati, Tamil, Telugu, Malayalam, Gurumukhi, Kannada, Assamese, and Urdu. Odia language is the most popular language used by 45 million peoples of Odisha. Many researches have been conducted to design an efficient Odia OCR. However, none of the work addresses the recognition of Odia conjuncts, whereas most of the works address recognition of basic characters

M. Nayak (✉) · A. K. Nayak
Department of Computer Science and Information Technology,
Siksha 'O' Anusandhan University, Bhubaneswar, India
e-mail: mamatanayak@soauniversity.ac.in

A. K. Nayak
e-mail: ajitnayak@soauniversity.ac.in

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_8

or numerals of Odia script [2, 3] only. Some of the characters of Odia script are made by combining two or more basic characters named as conjunct (also referred as compound) characters. Thus, leading challenge in design of an OCR for Odia script is to recognize those conjunct symbols.

Pattern recognition is the commonly used technique for character recognition. Depending upon the methods used for data analysis and classification, the pattern recognition models, used for character recognition, are broadly categorized into five different types: statistical model, structural model, template matching model, neural network-based model, and fuzzy-based model [4, 5]. Due to the learning and generalization ability of neural network-based model, it is commonly used by character recognition systems. Most of the ANN users rely on traditional schemes based on multilayer backpropagation network. Also, the accuracy in recognition of the characters is mostly influenced by the features extracted for each character. In this paper, we have used binary feature extraction technique and a novel optimized character recognition model by hybridization of differential evolution network with ANN.

The content of this paper is described as follows. Section 2 briefly describes the properties of conjunct characters in Odia script. Section 3 discusses the use of multilayer backpropagation neural network (MLBPN) and Sect. 4 presents the proposed differential evolutionary optimized backpropagation network (DEOBPN). The final section concludes with a comparison of the result obtained for both the approaches.

2 Properties of Conjunct Characters

In this section, we report various characteristics, peculiarities, and motivations of Odia script with an emphasis on conjunct characters. Most of the commercial OCR applications are concerned with the machine printed or handwritten, basic, and

Table 1 Conjunct characters of Odia script

Character	Construction	Character	Construction	Character	Construction	Character	Construction	Character	Construction
କ୍ଳ	ଡ + ୍ + କ	ଖ୍ଲ	ଖ + ୍ + ଚ	ଣ୍ଟ	ଣ + ୍ + ଟ	ନ୍ତ	ନ + ୍ + ତ	ମ୍ପ	ମ + ୍ + ପ
ଙ୍ଖ	ଡ + ୍ + ଖ	ଖ୍ଛ	ଖ + ୍ + ଛ	ଣ୍ଠ	ଣ + ୍ + ଠ	ନ୍ଥ	ନ + ୍ + ଥ	ମ୍ଫ	ମ + ୍ + ଫ
ଙ୍ଗ	ଡ + ୍ + ଗ	ଖ୍ଜ	ଖ + ୍ + ଜ	ଣ୍ଡ	ଣ + ୍ + ଡ	ନ୍ଦ	ନ + ୍ + ଦ	ମ୍ବ	ମ + ୍ + ବ
ଙ୍ଘ	ଡ + ୍ + ଘ	ଖ୍ଝ	ଖ + ୍ + ଝ	ଣ୍ଢ	ଣ + ୍ + ଢ	ନ୍ଧ	ନ + ୍ + ଧ	ମ୍ଭ	ମ + ୍ + ଭ

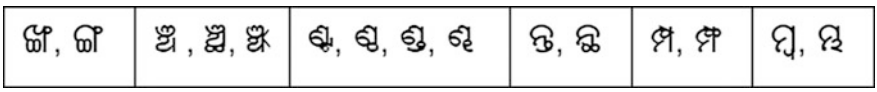


Fig. 1 Confusing conjunct character pairs

well-separated characters. As conjunct characters are also contained within the Odia language set, designing an OCR for Odia script to recognize the character, those that are formed by the combination of two basic symbols (characters) with the support of a joiner as shown in Table 1, is our main objective.

As shown in Fig. 1, we address that few compound/conjunct characters are identical with respect to their visualization, and thus they are also mentioned as confusing characters.

3 Multilayer Backpropagation Neural Network (MLBPN)

Backpropagation neural network (BPNN) is probably the most popular type of neural network [6–8] in which supervised learning algorithm is used to train for a given set of patterns with known classifications. Each entry of the sample set is presented as input to the network, and then the network generates its output responses of the sample input pattern. During the process of learning, the errors are subsequently backward propagated through the network to adjust the weights. Specifically, the process modifies the weights and thresholds of the network in an iterative way so that the trained network fits the training samples well. Figure 2a shows the architecture of the MLBPN model. The mathematical model of BPNN used in the work is represented in Eq. 1, in which O_j is the result of output neurons, w_{ij} denotes connection weights between ith input neuron to jth output neuron, x_i denotes the input variable i, f be the transitive function, and a_j represents the threshold of jth output neuron:

$$O_j = \sum_i f\left(w_{ij} * x_i + b_i\right) - a_j \tag{1}$$

$$O = [w_1, w_2, \ldots, w_D] \text{ and} f(x) = \frac{1}{1 + e^{-x}} \tag{2}$$

The gradient descent method is used as the adaptive learning function for the weight updation in BPNN, and its mathematical formulation is defined as follows:

$$w^{new} = \propto \times w^{old} - \eta \frac{\partial E}{\partial w}, \tag{3}$$

in which momentum coefficient is α, learning rate is η, and error gradient is $\frac{\partial E}{\partial w}$.

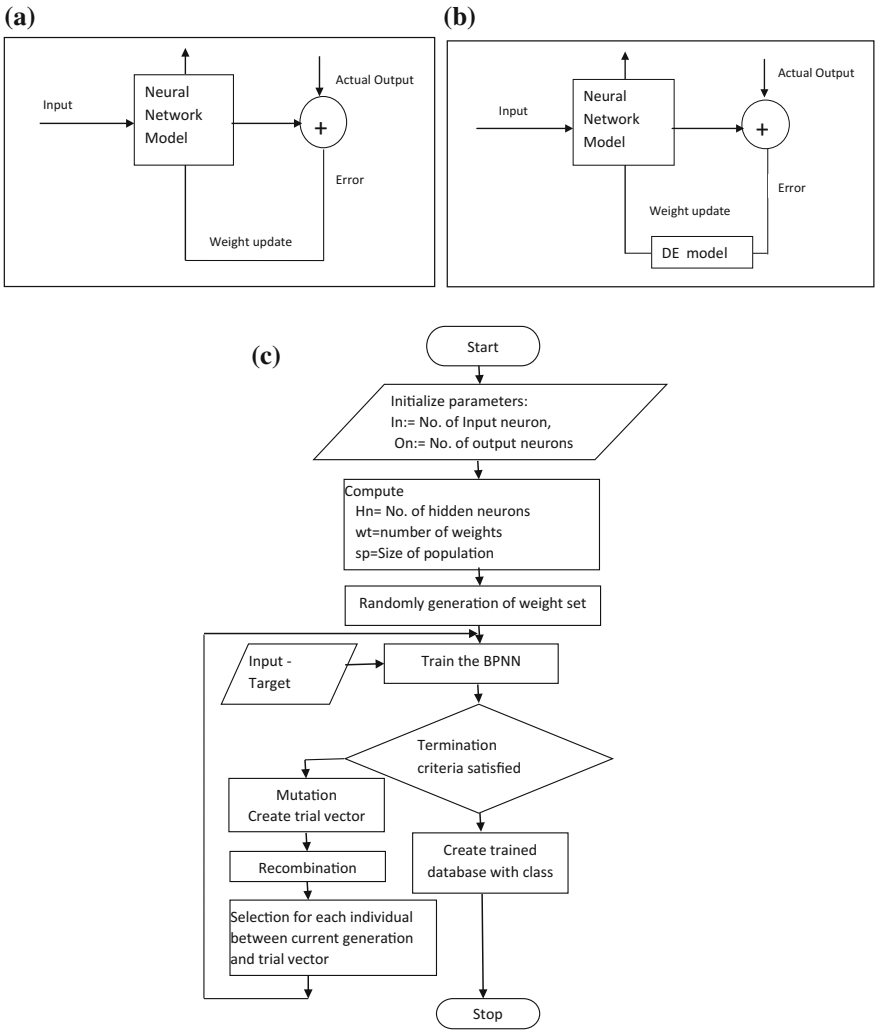


Fig. 2 **a**, **b** Block diagram of MLBPN and DEOBPN, **c** Flow diagram of DEOBPN

4 Differential Evolutionary Optimized Backpropagation Network (DEOBPN)

In this section, we propose the optimization of global search ability of BPNN using differential evolution (DE). DE algorithm is a population-based technique that allows each successive generation of solutions to evolve from the previous generation strengths [9, 10]. However, binary encoding is not supported by the DE, and also to self-adapt its parameters it does not use probability density function.

An important feature of DE algorithm is that during generation the step size depends on the difference vector ($x\text{-}_i$- $x\text{-}_k$) (where $x\text{-}_i$ and $x\text{-}_k$ are two individuals of the current population), crossover value, and distances between individuals [11]. The process of DEOBPN includes the population initialization, training BPNN, mutation, recombination, and selection as shown in Fig. 2b.

The network is trained for each training sample iteratively and finds the candidate solution of same population generation by generation. The number of connection weights used for the network is the parameters of a candidate solution. Each parameter represents a real value. The proposed DEOBPN model is presented in Fig. 2c. The population is said to be converged into two situations: either the percentage of error is less than a threshold value or after a specific number of iterations. The steps are as follows:

Step 1: Randomly, create a fixed size population of individual candidate solution such that each distinct x_i is consisting of k genes, whereas k represents the number of weight need to train the network. Each individual member x_i represents a point in the solution space. We assume that each jth gene of an individual x_i is assigned with values from −5 to +5.

Step 2: Train the neural network using the population and choose the individual as best solution for the population that having minimum value for the training error function defined as follows:

$$E = \sum_{i=1}^{I} (T_i - O_i)^2, \tag{4}$$

where I: total number of training inputs
T_i: target output of ith training vector
O_i: output generated for ith training vector

Step 3: (mutation/donor vector): For each vector x_i, randomly choose three other vectors. Create a donor vector by addition of difference of weights of both vectors to the third vector as follows:

$$v_{i,j} = x_{r_1,j} + F * \left(x_{r_2,j} - x_{r_3,j}\right), \tag{5}$$

where $F \epsilon (0, 2]$, $\mathrm{r1} \neq \mathrm{r2} \neq \mathrm{r3}$ and $\mathrm{r1}, \mathrm{r2}, \mathrm{r3}\, \epsilon$ population size.
The mutation factor (F) scales the difference of two vectors, we have used F as 0.5.

Step 4: (recombination/trial vector): Create trial vector u_i for next generation for all individual x_i by using binomially crossover operator with donor vector v_i.

The crossover operator is defined as follows:

$$u_{ij} = \begin{cases} v_{ij} \; if \; r \; and \; (i,j) \leq CR \\ x_{ij} if \; r \; and \; (i,j) > CR \end{cases} \tag{6}$$

(where CR ϵ [0, 1)).

The parameter CR denotes the crossover rate; we have used 0.9.

Step 5: (selection): Generate new population by comparing $E(u_i)$ with $E(x_i)$:

$$x_i = \begin{cases} u_i & if \; E(u_i) \leq E(x_i) \\ x_i & otherwise \end{cases} \tag{7}$$

Step 6: Repeat all steps until error percentage is less than a threshold value or the maximum number of iteration.

5 Data Used for DEOBPN

In this effort, 60 conjunct characters of three different font types of Odia script are considered as shown in Table 1, which are used for training and testing purpose. The most important aspect of character recognition system of any language is to analyze the input images, and then need to discover some special and distinct individualities of every symbol for discrimination of one symbol from other which is named as feature set/vector. We use the feature vector of every symbol by considering shape of obtained binary image. We follow the approach in which at first each image is transformed (i.e., scaled) to a size of 15 × 15, and then converted into a column vector with values considered in row-major order of size 225 × 1 [12]. In the same way, the feature values of all symbols those considered for training are being extracted and create a matrix that consists of 225 number of rows and 60 number of columns. The feature vector of one symbol is represented as a column of the generated feature matrix. As like training symbols, for all symbols used for testing, the feature matrix is computed. This method is simple, robust, and powerful.

6 Objective Function for Evaluation

Objective function is an assessment value that is used for recognition rate improvement as well as for the training error minimization. In this work, we use the objective function as minimization of the error (i.e., E) and is considered for each training input symbol, as defined as follows:

$$Objective\ function = Min\ (E) \tag{8}$$

In this paper, two different error functions are used to evaluate the performance defined as follows:

$$Mean\ Square\ Error_j = \sum_{i=1}^{I} (T_i - O_i)^2\ for\ j = 1, 2, 3, 4, \ldots, N \tag{9}$$

whereas T_i be the desire output of ith vector and O_i be the generated output:

$$Number\ of\ Character\ Mismatch = \sum_{i=1}^{GN} \left(N - N^{'}\right), \tag{10}$$

where N represents the total number of samples used for validation test and $N^{'}$ represents the total number of samples mismatched.

7 Result Analysis and Future Scope

The parameters set to train the MLBPN are described as follows: Momentum coefficient (α) is 0.8, learning rate (η) is 0.18, and threshold is 0.01. At first, the training of MLBPN is done for 5000 number of times. The error in each time of training is calculated with reference to Eqs. 9 and 10 and is backpropagated. Figure 3a shows the mean square error rate conversion and we have observed that, after 500 epochs, the curve is unvarying. From Fig. 3b, it is observed the number of mismatched characters at each iteration. Thus, we reach at a conclusion till end of the training; the percentage of error is not zero and few characters are not being trained properly from 60 character input.

Next, the proposed DEOBPN is trained for all characters. The parameters set to train the DEOBPN are as follows: population size (NP) is 200, each gene of an individual x_i is assigned with values from −5 to +5, mutation factor (F) is 0.5,

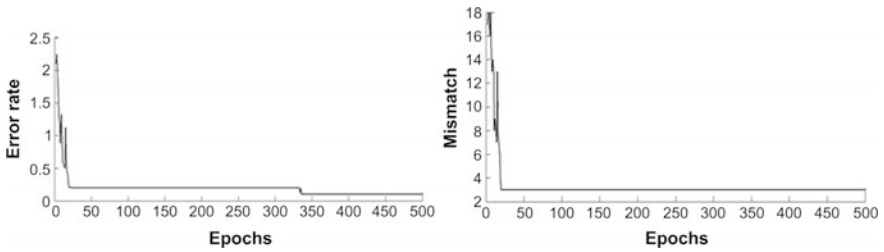


Fig. 3 (left) Rate of training error curves and (right) rate of number of character mismatches MLBPN

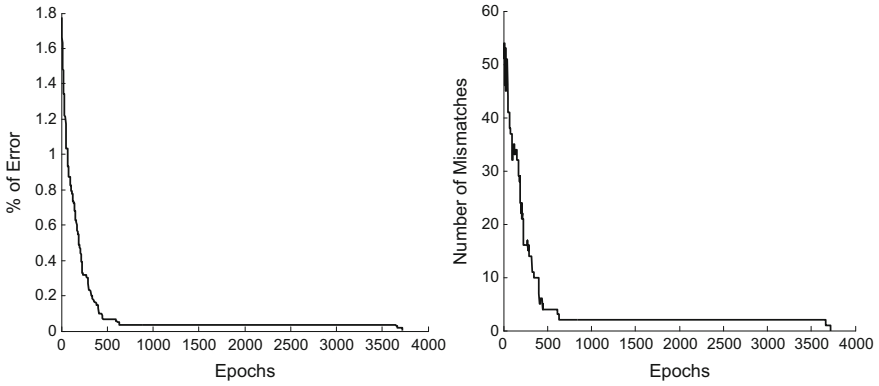


Fig. 4 (left) Training error rate curves and (right) rate of number of character mismatches of DEOBPN

and crossover rate (CR) is 0.9 chosen after a number of trials. Figure 4a, b shows the rate of error convergence by choosing mean square error and number of character mismatching the objective functions correspondingly. It is observed that the percentage of mean square error is convergence to zero in addition to the number of mismatch character which is also zero after 3700 number of iterations. Thus, the hybrid variant found (DEOBPN) provides proper solution with full reliability.

As a future direction of this work, the proposed model can be used for all characters of the Odia language; similarly by increasing the size of database, the recognition rate can be improved; likewise it can be further extended for hand-written document of Odia language and for different font types available in the Odia language.

References

1. Yadav, D., Sánchez-Cuadrado, S., Morato, J.: Optical character recognition for Hindi language using a neural-network approach. J. Inf. Process. Syst. **9**(1), 117–140 (2013)
2. Chaudhuri, B.B., Pal, U., Mitra, M.: Automatic recognition of printed Oriya script. Spec. Issue Sadhana Print. India **27**(1), 23–34 (2002)
3. Mohanty, S., Das Bebartta, H.N.: A novel approach for bilingual (English–Oriya) script. Identification and recognition in a printed document. Int. J. Image Process. (IJIP) **4**(2), 175–191 (2010)
4. Ghosh, D., Dube, T.A., Shivaprasad, A.P.: Script recognition—a review. IEEE Trans. Pattern Anal. Mach. Intell. **12**(2) (2010)
5. Bag, S., Harit, G., Bhowmick, P.: Recognition of Bangla compound characters using structural decomposition. Pattern Recogn. **47**, 1187–1201 (2014)
6. Barve, S.: Artificial neural network based on optical character recognition. Int. J. Eng. Res. Technol. (IJERT) **1**(4), 1–5 (2012)

7. Garris, M.D., Wilson, C.L., Blue, J.L.: Neural network-based systems for handprint OCR applications. IEEE Trans. Image Process. **7**(8), (1998)
8. Shrivastava, V., Sharma, N.: Artificial neural network based optical character recognition. Signal & Image Process. Int. J. (SIPIJ) **3**(5) (2012)
9. Das, N., Sarkar, R., Basu, S., Saha, P.K., Kundu, M., Nasipuri, M.: Handwritten Bangla character recognition using a softcomputing paradigm embedded in two pass approach. Pattern Recogn. **1**(48), 2054–2071 (2015)
10. Sarangi, P.P., Sahu, A., Panda, M.: A hybrid differential evolution and back-propagation algorithm for feedforward neural network training. Int. J. Comput. Appl. **84**(14) (2013)
11. Shamekhi, A.: An improved differential evolution optimization algorithm. IJRRAS **15**(2), 132–145 (2013)
12. Nayak, M., Nayak, A.K.: Odia-Conjunct character recognition using evolutionary algorithm. Asian J. Appl. Sci. **3**(4) (2015)

Email Classification Using Supervised Learning Algorithms

Akshay Bhadra, Saifuddin Hitawala, Ruchit Modi and Suraj Salunkhe

Abstract In the world of Internet today, huge amount of data is transferred between computers in the form of emails. Consequently, it is getting difficult to sort the important emails manually from the unimportant ones. Email classification has been extensively studied and researched in the past but most of the research has been in the field of spam detection and filtering. This paper focuses on classifying emails into custom folders that are relevant to the user. We have used two different approaches here—Naïve Bayes classifier and k-nearest neighbors algorithm. The Naïve Bayes classifier is based on a probabilistic model, while the k-nearest neighbors algorithm is based on a similarity measure with the training emails. We propose the method of using these two approaches in email classification, analyze the performance of these algorithms, and compare their results. Then, we propose some future work for further optimization and better efficiency.

Keywords Text classification · K-nearest neighbors algorithm
Naïve bayes classifier

A. Bhadra (✉) · S. Hitawala · R. Modi · S. Salunkhe
I.T. Department, Veermata Jijabai Technological Institute, Matunga,
Mumbai 400019, India
e-mail: bhadraakshay@gmail.com

S. Hitawala
e-mail: saifuddin.hitawala@gmail.com

R. Modi
e-mail: modi.ruchit6@gmail.com

S. Salunkhe
e-mail: salunkhesuraj10@gmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_9

1 Introduction

Emails have become an integral part of everyday life. Majority of computer users use emails to communicate with colleagues, e-businesses, families, and friends. Email messages in an inbox are sent, received, and accumulated in a repository over a period of time. Majority of the users never discard messages because the information contents might be useful later in time—for example, it might serve as a reminder of pending issues like payment of bills or might serve as a reminder for upcoming events. According to a survey [1], over 205 billion emails were exchanged in 2015 and the number is expected to rise to approximately 250 billion by the end of second decade of the twenty-first century. Hence, there is an immense need for classifying emails for better navigation and improving the efficiency of every user in the near future.

A lot of researchers have put their hard work in text classification using different algorithms and have suggested improvements in existing algorithms. Zhou et al. [2] experimented in text classification and the results have a good performance. They have used a modified KNN classifier which takes into consideration the classification when the number of training examples is not smooth. This problem is solved by clustering the preprocessed data and then classifying with a new KNN algorithm that makes run-time adjustments to the value of K.

Many techniques of supervised learning and unsupervised learning do exist in the literature for data classification. Somewhere between supervised and unsupervised learning lies semi-supervised learning. Here, some supervision is provided along with unlabeled data. Wajeed and Adilakshmi [3] applied semi-supervised classification of text documents to different types of vectors generated from those documents.

The major example of email classification occurs in classifying spam and non-spam emails. Harisinghaney et al. [4] compared the performance of three algorithms: KNN algorithm, Naive Bayes algorithm, and reverse DBSCAN algorithm based on four measuring factors, namely precision, sensitivity, specificity, and accuracy.

Chakrabarty and Roy [5] used minimum spanning tree method for clustering emails where the initial clusters (email folders) are unknown. The proposed algorithm yielded results that were more accurate than standard KNN and Naïve Bayes classifiers. Zhang et al. [6] used association features rather than primitive features and n-grams to improve the performance of Naïve Bayes classifier for text classification.

We have proposed email folders based on users' activities where incoming emails are identified and grouped into appropriate activities and related messages are grouped in the same activity.

Table 1 Sample work file schema

	Friend	In	Need	Is	A	Deed	Call
X_{M1}	1	0	0	0	1	0	1
X_{M2}	2	2	1	1	2	1	0

2 Preprocessing

2.1 The Bag of Words Approach

Before applying any machine learning algorithms, we have to preprocess the emails and represent them as a vector. Bag of words is a commonly used model in natural language processing (NLP). First step is the creation of work file—the collection of different words occurring in all the emails which are represented as comma-separated tuples representing each email. Let M1 and M2 be two emails in a training set:

- $M_{1:}$ "Call a friend."
- $M_{2:}$ "A friend in need, is a friend in deed." (Table 1)

Tokenization. Tokenization is a preprocessing method in which the email content is broken down into several tokens (words). These tokens are further used as input parameters for various NLP algorithms. Tokenization removes the punctuation from the sentences, separates them into words, and converts them into lowercase.

Stop Words Removal. Stop words are words encountered in the emails which do not provide any additional information (e.g., words such as "a", "in", "is", etc.). There are two ways in which stop words can be removed. First is by using a specific stop word dictionary, and another is by sorting all the words in all the emails by frequency count. The top n words in the stop list are then removed from the input emails.

Stemming. The process of transforming the words in sentences into its common base form is called stemming (e.g., "The crayons are different colors" is transformed after stemming to "The crayon be differ color"). The stemming library used here is Porter stemmer for python. All the above-mentioned methods of preprocessing have been applied using the stemming library in python. Finally, a "workfile.csv" is generated which has a frequency table of each word in every email. This step is necessary to eliminate the redundant words and in improving the accuracy of the applied algorithms.

3 Proposed Methods

3.1 Naïve Bayes Approach

The Naïve Bayes algorithm is a simple and efficient classification algorithm based on the assumption of the use of independent features for classifying new emails as stated in Bayes theorem. Features identified for emails are word frequencies which are

generally independent of each other for their weightage considered for classification purpose. So Naïve Bayes model turns to be beneficial in case of email classification. For example, a personal email might contain friendly and informal words like dear, oh, hi, etc., which can be used as independent properties having equal contribution in the calculation of probability for classifying the mail as personal.

Naive Bayes model is easier to develop and function on large datasets. It is also known to perform better in many complex scenarios as it needs small training data for understanding parameters needed for classification. Bayes theorem provides the method by which posterior probability P (c|x) can be calculated from P(c), P(x), and P(x|c). Look at the Eqs. (1) and (2):

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}. \tag{1}$$

$$P(c|x) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c). \tag{2}$$

- $P(c|x)$ is the posterior probability of *class* (c, *target*), i.e., category for given test email (x, *attributes*).
- $P(c)$ is the prior probability of *class, i.e., category.*
- $P(x|c)$ is the likelihood which is the probability of *test email* given *class.*
- $P(x)$ is the prior probability of test email.

Constructing classifier by Eqs. (1) and (2),

$$\widehat{y} = \underset{k\varepsilon\{1,\ldots,K\}}{argmax} p(C_k) \prod_{i=1}^{n} p(x_i|C_k). \tag{3}$$

Here "y" is the assigned class value to the test email depending on the class giving highest probability for the test email. "n" indicates the number of features of email considered for building feature set for email dataset.

Calculation of probability using Gaussian model:

$$p(x = v|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{(v-\mu_c)^2}{2\sigma_c^2}}. \tag{4}$$

Here, x is the feature of test email considered and c is the category label considered. μ_c is the mean of the attribute values in x, i.e., feature associated with category c and let σ_c^2 be the variance of the values in x associated with category c.

Results. We have considered a dataset of 1500 mails and have classified them into four categories. We have analyzed the working of Naïve Bayes by varying the ratio of mails considered for building training set so as to classify the test set data. We have built a feature set which composes words with a frequency of greater than 10 in the whole dataset after removing useless words from the set. The graph of the result is shown in Fig. 1.

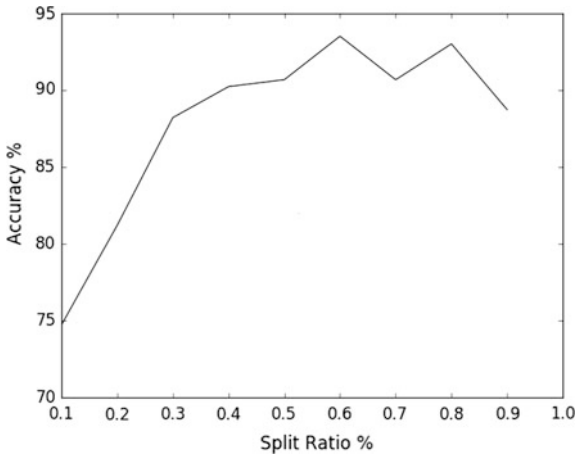


Fig. 1 The graph shows that 0.6 split ratio is the cutoff for best accuracy (approx. 94%) result with Naïve Bayes and after that point graph fluctuates

Analysis.

1. Time Complexity:

It is very efficient as it is linearly proportional to the time required to read in all the emails in the dataset, i.e., $O(|E|*L_{avg})$ where E is the count of number of emails in the dataset and L_{avg} is the average length of an email in the whole dataset.

2. Pros:

- It is easy and fast for even large datasets. It takes ~10 s for 1500 mails.
- Taking assumption of independence between features of email, a Naive Bayes classifier performs better compared to other models like logistic regression. Accuracy is good even if the training set is small compared to train set as we can see in the above graph for 0.1 split ratio, accuracy comes to be around 75%.

3. Cons:

- It performs well with input variables related to categories as compared to numerical variable(s). But features considered in emails are mostly numerical variables, so category advantage is not present for email classification.
- Limitation of Naive Bayes is the assumption of independent predictors. But in reality, it is almost impossible that we get a set of features in email which are completely independent as a pair of words can help in prediction of class compared to a single word as a feature.

3.2 K-Nearest Neighbors Approach

The k-nearest neighbors algorithm tries to find k emails in the training set that is closest to the test set. The class of the test set is then decided based on the majority

class of the k neighbors. To find these neighbors, we need a similarity measure. The similarity measure that we used is the feature vector of the bag of words approach in the preprocessing. The algorithm is as follows:

1. Divide the dataset created using the preprocessing steps into two sets—Training set and test set—with a split ratio of 0.67. Thus 67% of the total emails are used as training set while the rest are used as test set.
2. Choose a value of k or loop for different values of k.
3. For each email in test set,
   (a) Calculate the distance* of the email with all the emails (neighbors) of the training set.
   (b) Sort the emails according to the distance and find k-nearest neighbors.
   (c) Assign that class to the test email that is most common among its k-nearest neighbors.
4. Repeat step 3 until all test emails are assigned their classes.

*distance: The distance is calculated using the Euclidean distance formula. We have two feature vectors, one of the test emails say X and other of the training email say Y. These two vectors are nothing but the tuples in the work file created in preprocessing. The column values of these two vectors are x_1, $x_2 \ldots x_n$ and y_1, $y_2 \ldots y_n$, respectively, where n is the number of columns in the tuples. The distance is then calculated using Eq. (5):

$$distance = \sqrt[2]{\sum_{i=1}^{n} (x_i - y_i)^2}. \tag{5}$$

This distance is smaller between emails that belong to the same category and larger between emails belonging to different categories. This property helps in identifying the class of the emails.

Results. We applied the KNN algorithm on a dataset of 1500 emails which has four categories of emails, namely "calendar", "meeting", "personal", and "recruiting". This dataset was splitted into training and test sets depending on the split ratio. We also executed the algorithm for different values of k to find the best possible value of k. Once the test emails were classified, we calculated the accuracy based on their actual classes (categories). Figure 2 shows the graph of accuracy versus the value of k.

Analysis. As we can see the accuracy fluctuates reaching a maximum at $k = 25$ after which the accuracy decreases. This is due to the fact that the number of mails in the calendar category is more as compared to that of other categories. Thus, the features pertaining to the calendar category, i.e., today, time, week, etc., overpower all the other features, thus classifying more mails to the calendar category and decreasing the accuracy.

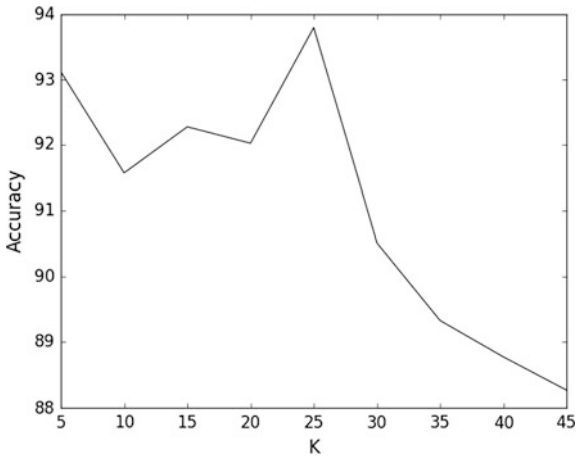


Fig. 2 Graph of accuracy versus values of K in KNN algorithm

3.3 Using K-Nearest Neighbors for Classification of New Emails

In the above approach, we have used KNN algorithm on the given dataset to prepare a dictionary and work file, and then to divide the dataset into training and test sets to classify the test set using the training set. Here, we discuss how we can use the already prepared work file as a training set to classify new emails that will pop up in the user's inbox.

The algorithm is as follows:

1. Prepare the work file of a labeled dataset using the extract.py and load it as a training set, keeping a list of all the words of the training set in training words list.
2. Read the new emails and prepare a words dictionary with word count.
3. Prepare the test set as follows:
   (a) For each new email:
      (i) For each word in training words list, if the word in training words list is also in new email, add the word count to test set.
      (ii) Else add 0 to test set.
4. Use the test set (step 3) with the training set (step 1) in the KNN algorithm to classify the new emails.

Results. This algorithm was run on a training set of ~7000 emails. These emails were labeled into the following four categories: "calendar", "meeting", "personal", and "recruiting". Five new emails of each category were then fed into the classifier and the results are shown in Table 2. Figure 3 shows that the graph is very stable

Table 2 Classification of new emails using KNN algorithm

K	Calendar (0)					Meeting (1)					Personal (2)					Recruiting (3)				
1	0	0	0	0	0	1	1	1	1	1	0	2	2	2	2	0	2	0	0	0
3	0	0	0	0	0	1	1	1	1	1	0	2	2	2	2	0	3	0	0	0
7	0	0	0	0	0	1	1	1	1	1	0	2	2	2	2	0	3	3	0	0
15	0	0	0	0	0	1	1	1	1	1	0	2	2	2	2	0	3	2	0	0
24	0	0	0	0	0	1	1	1	1	1	0	2	2	2	2	0	3	2	0	0
33	0	0	0	0	0	1	1	1	1	1	0	2	2	2	2	0	2	2	0	0
42	0	0	0	0	0	1	1	1	1	1	0	2	2	2	2	0	3	2	0	0
50	0	0	0	0	0	1	1	1	1	1	0	2	2	2	2	0	3	2	0	0

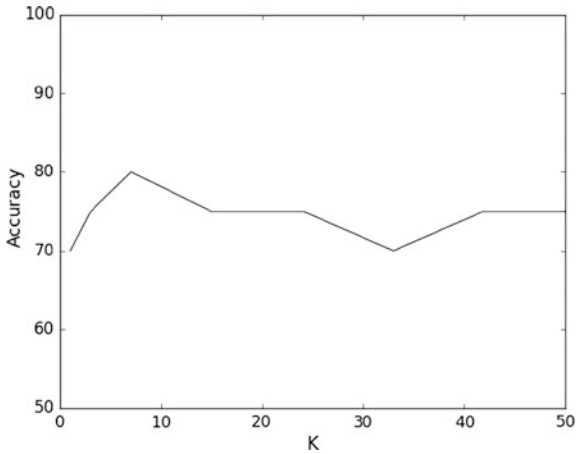


Fig. 3 A graph of accuracy versus values of k

and the accuracy is maximum when k = 7. Also, the accuracy variance is very less and almost constant around 85%.

Analysis. The algorithm classifies new emails using the KNN algorithm with a slight difference in the way the test set is created. Only the words that are in the training set and in the new emails are counted and included in the test set. This works very well for categories that always have few words in their emails, like the meeting category will always words like meeting, conference, delegates, etc. in its emails. It can be seen that the accuracy decreases as the value of k increases. From Table 2, it is observed that majority of the emails from recruiting category are not classified correctly.

The reason is that there were only few training samples of recruiting category (735) as compared to the emails of other categories (2000 + in each). Because of this, when the value of k is greater, emails of other categories start becoming neighbors of the new email and the new email gets classified incorrectly. Therefore, it is better to have a training set with equal number of sufficient emails in each category. It should also have a variety of emails in each category and the value of k should be chosen accordingly.

4 Comparison of Algorithms

As we can see from the algorithms implemented above, Naïve Bayes algorithm gives a better accuracy as compared to k-nearest neighbors algorithm. This is due to the fact that KNN algorithm compares each test mail with its neighbors (training mails) which may or may not be accurate. On the other hand, Naïve Bayes algorithm works on a probability function which is based on the model created. Thus, there is no need for physical comparison of mails as in KNN. Moreover, Naïve Bayes works in seconds, whereas KNN requires minutes to execute each test mail for each value of k. However, after extracting only the relevant features using a cutoff value, the time taken to run one epoch is comparable to that of Naïve Bayes. Moreover, the algorithm has been optimized by providing weights to the subject and body of the mail, i.e., 4:1 (number of folders: 1); there is a significant improvement in the accuracy as well as the time complexity. Thus, summarizing, both the algorithms have their own pros and cons. The algorithm to be used depends upon which one gives better results for a given dataset.

5 Conclusion and Future Work

We have seen how our proposed algorithms help classifying the emails into supervised categories. There is a lot of work that can be done on these algorithms like using parallel processing algorithms for concurrent classification of incoming emails. These algorithms also take a lot of time to compute the result, and hence methods to reduce it can be explored.

So far we have considered each and every word separately. We can further analyze the combination of sensible words together in the form of di-grams, tri-grams, and n-grams to improve the accuracy of classification.

Acknowledgements We would like to thank Prof. V.K. Sambhe for providing guidance to us in this project, giving important suggestions and helping in carefully reviewing this paper. We would also like to thank the other faculty members of the Computer Engineering and Information Technology Department of V.J.T.I. for their valuable inputs and suggestions.

References

1. Email Statistics Report, 2015–2019 conducted by, The Radicati Group, Inc. A Technology Market Research Firm, Palo Alto, CA, USA
2. Zhou, L., Wang, L., Ge, X., Shi, Q.: A clustering-Based KNN improved algorithm CLKNN for text classification. In: 2010 2nd International Asia Conference on Informatics in Control, Automation and Robotics (CAR), vol.3, no., pp. 212–215, 6–7 Mar 2010

3. Wajeed, M.A., Adilakshmi, T.: Semi-supervised text classification using enhanced KNN algorithm. In: 2011 World Congress on Information and Communication Technologies (WICT), vol., no., pp. 138–142, 11–14 Dec 2011
4. Harisinghaney, A.; Dixit, A.; Gupta, S.; Arora, A., "Text and image based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm," in Optimization, Reliabilty, and Information Technology (ICROIT), 2014 International Conference on, vol., no., pp. 153–155, 6–8 Feb. 2014
5. Chakrabarty, A., Roy, S.: An optimized k-NN classifier based on minimum spanning tree for email filtering. In: 2014 2nd International Conference on Business and Information Management (ICBIM), vol., no., pp. 47–52, 9–11 Jan 2014
6. Zhang, Y., Lijun, Z., Jianfeng, Y., Zhanhuai, L.: Using association features to enhance the performance of Naive Bayes text classifier. In: 2003. ICCIMA 2003. Proceedings of the Fifth International Conference on Computational Intelligence and Multimedia Applications, vol., no., pp. 336–341, 27–30 Sept 2003

Multilayer Perceptron Neural Network Based Immersive VR System for Cognitive Computer Gaming

P. S. Jagadeesh Kumar

Abstract Culmination in the simulation of immersive virtual reality system has provoked intensifying stratum of common sense and splinter group in inherent character that closely mandrill corporeal authenticity in building efficient computer game using multilayer perceptron neural network. In this paper, the winch in motion supremacy through two-level formation steering is conferred and kinematic movement algorithm heralds an intent momentum; subsequently, multi-core processing and hyper-threading are dilapidated to put its direction to assist in the formation of 3D video games. Nevertheless, immersive virtual reality system bequeaths computer-stimulated behaviour and dynamism of their individual for efficient swiftness by means of agent-based artificial intelligence.

Keywords Computer gaming · Immersive virtual reality · Kinematic movement algorithm · Multilayer perceptron neural network · Multi-core processing and hyper-threading · Two-level formation steering

1 Introduction

With virtual reality knowledge, onlookers can make a distinction amid aloofness and spatial associations involving dissimilar item constituents more rationally and precisely than with conservative apparition tools. Computer simulation is neurotic with building succession of metaphors that when exhibited uninterruptedly, at rightfully elevated swiftness, bestow the hallucination of intelligible method of the image rousing in definite comportment. It is plausible to consign provisions on computer gaming in the course of conformist advance to gather together these rations to exploit proficient performance in virtual reality system [1]. The contribution of the exceptionally educated designer and developer may surpass what

P. S. Jagadeesh Kumar (✉)
Department of Computer Science and Engineering, Marathwada Institute of Technology, Aurangabad, Maharashtra, India
e-mail: dr.psjkumar@yahoo.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_10

might perhaps be pragmatic by technology. In distinction, not all individuals are talented enough in building high-quality facets of animations in video gaming which has the occasion, stamina, competence, or principal to accomplish accordingly. Additionally, for crystal-clear brand of content, like computer games, human involvement in the real-time profligacy of requirements might not be practicable. Consequently, computer gaming endeavours to come up with schemes by the chase of trending technologies that flock the specified provisions which might be realistic by means of artificial intelligence.

2 Artificial Intelligence and Computer Gaming

Artificial intelligence (AI) is a training, which make computers to attain the thinking ability that human being and animals are proficient enough [2]. The development needs to be consigned prior to recuperate into AI coding is illustrated in Fig. 1. The entity practices can then be slit in. There are methods that once enveloped can work individually, and all the algorithms are reasonably self-regulating. For demonstration, it might be adequate to just employ the performance. The system kowtows to a typical constitution for AI performances: it can be agreed upon execution time, it obtains data from a vital messaging proposal, and yet it develops in a normal configure. The fussy set of crossing points utilized proves its own advancement prejudice. They were intended to be quite easy; consequently, the algorithms are not imposed by message system. By the similar voucher, there are simple optimizations that can be dappled and comprehended yet again for lucidity. The complete AI scheme has an analogous crossing point to the cryptogram with abundant alacrity and reminiscence restraints. Other AI engines on the souk have a dissimilar configuration, and the explicit engine will probably put extra constrictions on realization. The high-quality AI structure facilitates salvage, restoring and expansion time [3].

2.1 Agent-Based Artificial Intelligence

In this perspective, agent-based AI is a method, generating monarch characters that acquire statistics from the game data and decide what events to get supported by the statistics, and cart out those events. It is able to be seen as bottom-up intend; it starts by functioning out how every disposition will perform by executing the AI that is preferred to prop up. The overall performance of the complete game is merely a gathering of how the personalities' mannerism labours in concert. The first two rudiments of the AI represent the faction and resolution of the AI for a mediator in the game. In distinct, a non-agent predestined AI requests to labour how the whole thing must ensue from the top-down and formulates a distinct scheme to imitate everything.

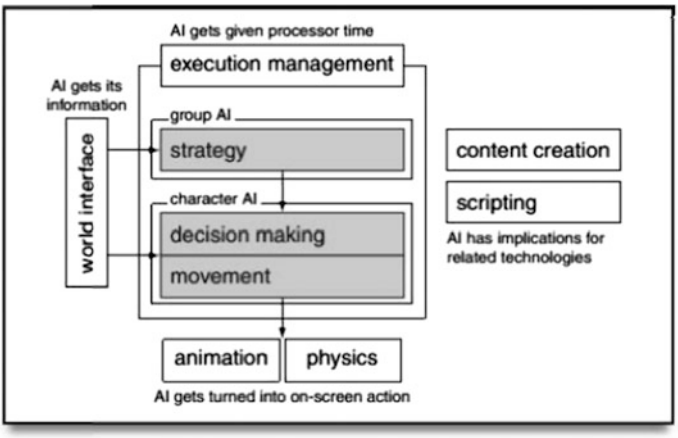


Fig. 1 AI model for game design

2.2 *Multi-core Processing and Hyper-Threading*

Recent processors have more than a few implementation trails lively at the similar instance. Code is conceded into the processor, isolating into numerous conduits which accomplish simultaneously. The consequences from every pipeline are then reunited into the concluding effect of the unique code. When the outcome of one conduit depends on the outcome of any more, this can engross backtracking and replicating a set of training. The couple of algorithms on the processor that efforts away how and where to rip the structure foresees the probable effect of persuaded reliant contrives, which is labelled as branch prophecy. This device of processor is termed as super-scalar. Regular threading is the procedure of permitting unusual bits of code to practice at similar instance. Hyper-threading is an Intel brand for exercising the super-scalar temperament of the processor to propel dissimilar threads along unlike conduits. Every pipeline can be provided a diverse thread to practice, allocating threads to be indisputably coursed in parallel. Hyper-threading is accessible barely on definite processors and operating systems [4].

2.3 *Artificial Intelligence Schematic*

The definitive arrangement of the AI engine may gaze somewhat as shown in Fig. 2. Data is fashioned in a contrivance, which is afterward enclosed for exercise in the game. While a stage is laden, the game AI actions are shaped from intensive data and chronicled among the AI engine. Throughout game participate, the major game code entitles the AI engine which revises the actions, receiving data from the global boundary and finally relating their productivity to the game information.

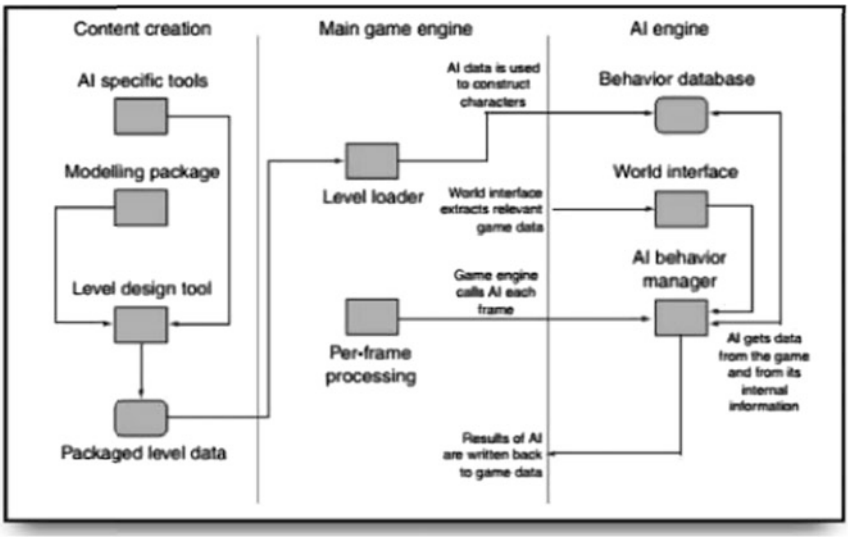


Fig. 2 Artificial intelligence schematic

The techniques employed depend a lot on the genus of the game corporeal. As the game is peripheral, it is necessary to take a castoff about the methodology to excavate the actions [5].

2.4 Kinematic Movement Algorithm

Kinematic movement algorithms utilize static information and yield the preferred rapidity. The outcome is frequently just an on or off and an objective track, poignant at complete velocity or being motionless. Kinematic procedures do not employ stepping up, though the sudden amends in speed might be curved above quite a lot of frames. Numerous games make simpler stuffs still advance and oblige the compass reading of a personality to be in the course it is moving. If the personality is motionless, it features whichever a predetermined path or the previous path it was pitiful. If its movement algorithm precedes an objective speed, afterward it is worn to set its direction [6].

2.5 Two-Level Formation Steering

Strict geometric be capable of collective in configurations with the suppleness of an embryonic advance by means of a two-level steering scheme. Geometric formation

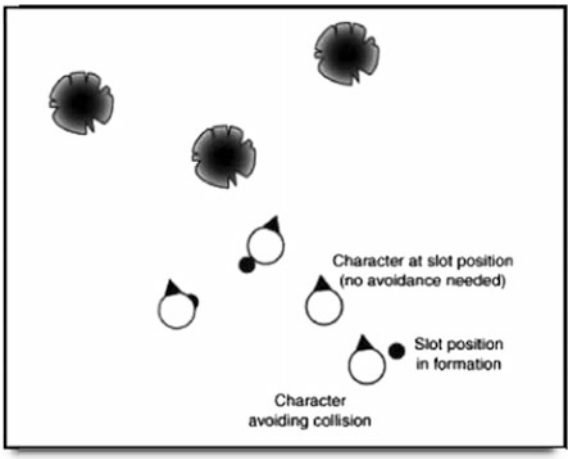


Fig. 3 Two-level formation motion

is definite with a permanent prototype of slits, immediately as prior to. Relatively inserting every personality in its slit, it trails the sprouting method by employing the slot at objective position for a disembark performance. Personalities can have their own fender-bender dodging manners and another complex routing requisite. It is a two-level steering since there accomplished two steering structures in succession: primarily the head steers the construction prototype, and after that every disposition in the structure steers to continue in the outline. Providing the head does not budge at utmost rapidity, every quality will have a number of suppleness to languish in its niche while intriguing the report of its surroundings. Figure 3 illustrates the number of agents exciting in V shape all the way through woodland. The attribute V silhouette is perceptible, other than every character has enthused unconscientiously from its slit location to evade walloping into foliage. The slit that a personality is annoying to attain might be momentarily unfeasible to accomplish, but its routing algorithm guarantees to perform logically [7].

2.6 Multilayer Perceptron Neural Network Architecture

A range of learning systems is likely for the personality. So far, the mainstream of labels depends on neural networks; fortification learning would moreover be a reasonable system to endeavour. On behalf of a neural network learning algorithm, the present is a combination of two varieties of regulation: strong regulation from examination and weak regulation from player response [8, 9]. As an array of unusual network, structural designs be able to exploit for this kind of game, a multilayer perceptron network is subsisted, as shown in Fig. 4. The input for the neural network obtains the background data from the game. The output layer includes the nodes calculating the form of deed and the entity and circuitous point

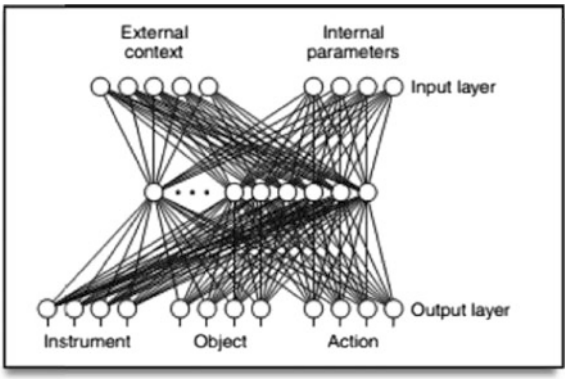


Fig. 4 Multilayer perceptron neural network architecture for 3D game design

of the exploit. In autonomous learning, the network can be secondhand to build assessments for the temperament by providing the present background as an input and interoperating the action from the production.

3 Pragmatism and Virtual Reality

Pragmatism is vital in the contraption of virtual reality possessions when dynamic personality inhabits the analogous department as the live player in the live contiguous. In such state of affairs, implicit personality exploits must illustrate to set out with the actions and factions in the live contiguous or the absurdity will be blatant to the observer, for example, a personality that is soaring and thus being suggested to the fly ought to visually stint an article that performs be factual flying in the similar panorama in the analogous significance [10]. Alternatively, the habitual implements for stylish 3D character imitation do not integrate influential information, which denote that the verve of character association can survive totally in the animation gallows. Numerous animation methods assimilate corporal reproduction of rigid and agile firms, soggy, chatters and ethical fibres.

3.1 Virtual Reality Technology

Virtual Reality (VR) is a computer-generated, 3D milieu in which a plurality of human accomplices, fittingly edged, may employ and manoeuvre replicated corporeal rudiments in the atmosphere and, in a number of structures, might connect and interrelate with depictions of other individuals, precedent, recent or illusory, or with imaginary mortal. It is a computer-based skill for imitating visual-acoustic and other sensory features of versatile contiguous. VR integrates 3D expertise that provides a real-life delusion. It crafts a model of real-life conditions. Thus, VR

submits to an immersive, cooperative, vision centred, 3D computer fashioned setting and the blend of knowledge requisite structure such an atmosphere [11]. By engrossing observers in computer-created stereoscopic surroundings, VR knowledge cracks down blockades among persons and workstations. This technology reproduces natural stereoscopic analyzing methods by employing computer skill to make right ogle and left ogle images of a specified 3D entity or panorama. The spectator's intelligence incorporates the data from these two perceptions to generate the insight of 3D liberty. Therefore, VR tools generate the fantasy that displayed articles have deepness and occurrence ahead of the smooth image predicted onto the panel.

3.2 Controlling Actions and Behaviour

High-intensity control practices construct it probable to provide activities to mainframe produced personality that formulate them to appear intellect; that is, they interrelate with other dispositions with analogous chattels and act in response to ecological conditions in a significant and productive way [12]. Such circumstances have the forthcoming of receiving libretto statistics as put in and produce computer-made successions as production. Application vicinity includes fabrication animation and interrogative computer games. Furthermore, investigators are now scrutinizing customs of having virtual persons to execute composite responsibilities constantly.

3.3 VR-Based 3D Game Design

VR has slash athwart to every single aspect of human stab mechanized/commerce, inquiry, safety, vacation activities and medicine amongst others. The exhilarating ground of VR has the latent to adapt the lives in numerous behaviours. There is countless relevance of VR at the moment and at hand will be lots more in the upcoming. A lot of VR functions have been urbanized for industrializing, edification, reproduction, plan valuation, architectural, ergonomic, recreation of assembly series and continuation missions, support for the physically challenged, revision and healing of phobias, amusement, quick archetype and to a large extent. VR expertise is at present extensively recognized as a foremost infiltrate in the technical progress of science [13].

4 Game Design and Development

Playing computer games is a cool leisure commotion for youthful populace. Not startlingly, a lot of these fanatic visions will one day enlarge computer games. This witnessed the utilization of game devise as a medium to educate adolescent computer science. Designing computer games includes a lot of features of calculation, counting computer realistic, artificial intelligence, sanctuary, dispersed encoding, imitation and software. Game improvement also fetches into co-operate part of the broadminded sculpture, science and psychology. Construct a modern viable computer game is an extremely complicated assignment that classically entails a mammoth budget and an increased team.

4.1 Design Challenges

At the present time, game construction pretends a number of implausible confronts. Panels come together to construct a distinct game, and then break up. Preparations of instants on game engine enhance gradually. Code components, built for exact games, proffer take away 30% reclaim in consequent labours. Firearms are presently the games that have reusable trainings. These engines are transitory, though, secondhand for a small number of games only, accredited at inflated outlay, charge additional and seize further years to build up. Game construction time augments progressively as graphics cards offer additional visual ability and the players' stipulate more verisimilitude. The command for superior personality and tale rises with the intricacy of visual exhibit with liberate to both novel and additional multifarious than constant gaming. Game oblige modernism is flattering an aggressive requisite. On behalf of the game diligence prolongs to cultivate, further genus has to befall more complicated, with improved back narrative and methodically investigated, urbanized and deployed introductory gameplay knowledge. The players are flattering still more captivated of convenient amusement proposals, and calculating alacrity is fetching increasingly quicker [14]. Furthermore, the method can perfunctorily adapt the animatronics to account for the difference amid the pioneering simulation and actually suitable vibrancy.

4.2 Building the Game Environment

Indulgencing and tormenting will contribute a key role in the games research schema. When canvassers situate persons in significant single-player components, they ought to decide what occurs throughout gameplay and it means the practice influences the players. Existing serious game training to great extent simulates the need for characters behavior to afford interacted environment. When the game

finishes, the human supervise informs the canvassers which player succeeded and why? To purify this procedure, designers have to obtain a computerized perceptive and investigation potential that can fabricate a high-level description of what ensued throughout gameplay above a particular stage, from a careful viewpoint, with the alternative to inquiry the scheme for supplementary thorough statistics. Frequent protection, sanctuary and didactic functions necessitate such computerized learning if gaming is to craft significant donations to the severe game area. In accumulation, the amusement engineering may discover such study helpful as an advertising and game-sophistication device. Pedagogy and chronicle assimilation engage decisive theory and budding performs for introducing knowledge occasion into tale, such that contestants discover the anecdote immersive and amusing since the entrenched tutoring remnants subsidiary to it. The game engineering has by now observed the breakdown of edutainment, a discomfited assortment of edifying software flippantly speckled with fixture interface and attractive dialogue. This malfunction demonstrates that fairy tale must approach first and investigate spotlight on coalesce teaching with narrative conception and the game improvement course.

4.3 Implementation of Immersive VR System

An immersive VR scheme is the main straight practice of virtual atmosphere. At this time, the consumer also bears a head-mounted display or employ a number of forms of head-joined exhibit such as binocular omni-orientation monitor to sight the virtual surroundings, in count to several trailing diplomacies and herpetic strategies. It utilizes diminutive scrutinize located in obverse of each ogle which affords stereo-system, bi-optical, or mono-optical imagery. Fashioning expertise engages the game thespian's mind through sensory inspiration and provides technique for escalating the intelligence of occurrence donated to edifying an emotion of fascination. This labour comprises computer animatronics resonance and hepatices, sentimental compute logic on human condition of passion, and superior consumer crossing points. Sensory means of phrase examines theatres an essential task in games paraphernalia of an immersive virtual reality-based waiting hall as shown in Fig. 5.

4.4 Realization of Artificial Intelligence Ecosystem

Classically, many mortals are present in a group game, and they are the connections of all genera that construct the game world attractive for the thespian. As a group, it pays plenty of space for attractive approach: one class can be worked to manipulate another, which can guide to unforeseen resolution to enigmas in the pastime. Using multilayer perceptron neural network, the genus can be approved, where the thespian frequently is tasked with susceptible group of persons in a waiting hall

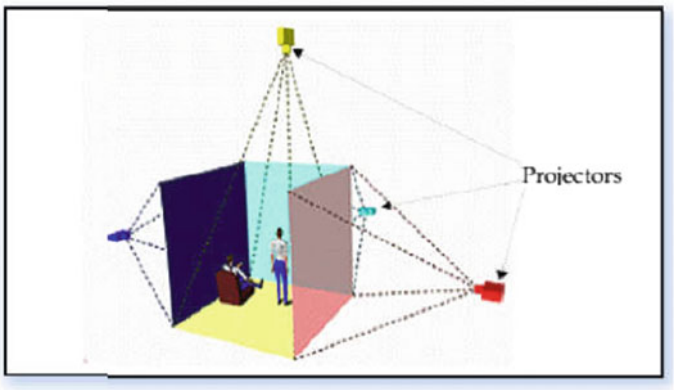


Fig. 5 Schematic representation of a waiting hall

Fig. 6 AI-based immersive VR system of a waiting hall

through the schematic representation of an immersive virtual reality system as shown in Fig. 6. When designing the ecosystem of a game, superfluous, and optimistic but unforeseen, possessions can be commenced. To prevaricate a render down in the game echelon, the entire individuals are quickly chomped, and some indispensable strategies necessitate to be pursued. The most significant fraction of receiving a playable game is to construct and pinch typescript. The conductional animation constitution that chains appropriate high-level communication and the path of self-governing implicit agents is urbanized by the following:

1. Facilitating an agent to stand for and trust in a manner that is instinctive to the animator through multi-core processing and hyper-threading;
2. Permitting a user to describe the agent's illustration in provisos of events and their property with the help of kinematic movement algorithm;
3. Fusing the reparation of a high-level reckoning scheme with an inferior level in discreet demeanour system using two-level formation steering;

4. Authorizing an interpretation agent to intellect, so that it is able to situate in a science-based lively world exploiting immersive VR system;
5. Allowing an agent to obtain guidance on how to perform in the way of a draft sketch, the particulars of which the mediator can involuntarily surmise;
6. Subsisting to express an agent exclusive of sacrificing its independence at its execution by means of multilayer perceptron neural network.

5 Conclusion

In sophisticated cognitive 3D computer gaming explore, cramming implicit sovereign mediators are placed in vibrant virtual worlds. Previously, canvassers in computer game design have exploited techniques from control theory and arithmetical normalization to set aside them to undertake the low-level control problem. To ensure in self-motivated implicit worlds, a hypothesis of deed that permits virtual reality-based independent personality to identify the cause is practical. An amazingly functional scaffold for high-level control that coalesces the profits of speedy and rationale scheme is realized. It permits expediently identifying the high-level performance of the intrinsic computer game personality to display. A great deal of the premature labour in computer gaming includes building competent achievements of completely superior technical hypothesis, or falling that, estimation in contributing presumption that make tranquil level-headed gazing fallouts. It was not protracted prior to intricate vague tribulations, such as the control quandary for virtual reality-based video games. Intended for a lot of features, the majorities are conspicuous cognitive procedures, systematic speculation is a lot less absolute, and computer animatronics is supposed to partake an essential position, together as a trial and a dynamic power, in rising innovative thoughts through artificial intelligence-based immersive virtual reality system for cognitive 3D video gaming.

References

1. Daniel, K.: Ethics of Virtual Reality Applications in Computer Game Production. Philosophies, 1, pp. 73–86 (2016)
2. Firas, S., Raphael, F., Damien, E.: Artificial intelligence in video games: towards a unified framework. Int. J. Comput. Games Technol. (2015) https://dx.doi.org/10.1155/2015/271296
3. Eva, P., Stephen, P., McGuire, A.: Using virtual reality and videogames for traumatic brain injury rehabilitation: a structured literature review. Games Health J. **3**(4), 202–214 (2014)
4. Wei Ming, Y.: New Integration Technology for Video Virtual reality. ACIIDS, Part 1, pp. 111–117, Springer, Berlin Heidelberg (2012)
5. Himma, K., Ed., Tavani, H.: Virtual Reality and Computer Simulation. Handbook of Information and Computer Ethics, Wiley (2008)

6. Mark, O.R.: Interactive narrative: a novel application of artificial intelligence for computer games. Assoc. Advanc. Artific. Intelligen. (2012)
7. Joseph, J.L., Jr.: Bringing VR and spatial 3D interaction to the masses through video games. IEEE Comput. Soc. (2008)
8. Zyda, M.: From visual simulation to virtual reality to games. IEEE Comput. Soc. 25–32 (2005)
9. Anderson, E.F.: Playing Smart—Artificial Intelligence in Computer Games. Conference on Game Development (2003)
10. Siong, G.: Low.: Understanding Realism in Computer Games through Phenomenology. CS378 Spring Term Paper (2001)
11. Simon, M.L., Michael, M., Mike, P.: Artificial and computational intelligence in games. Artific. Comput. Intelligen Games. **2**(5) (2012)
12. Ian, M.: Artificial Intelligence for Games. Elsevier, Morgan Koufmann Publishers, San Francisco (2006)
13. Dustin, F., Otmar, H., Abigail, S.: The Role of Physical Controllers in Motion Video Gaming. DIS, June 11–15, 2012, Newcastle, UK, ACM (2012)
14. Aliza, G.: Academic AI and video games: a case study of incorporating innovative academic research into a video game prototype. In: IEEE Symposium on Computational Intelligence and Games (2005)

Computer-Aided Therapeutic of Alzheimer's Disease Eulogizing Pattern Classification and Deep Learning Protruded on Tree-Based Learning Method

P. S. Jagadeesh Kumar

Abstract Alzheimer's disease, the prevalence genre of non-curative treatment, is probable to rumble in the impending time. The ailment is fiscally very lavish, with a feebly implicit cause. Premature therapeutic of Alzheimer's disease is extremely imperative and thus a titanic covenant of deliberation in the growth of novel techniques for prior discovery of the illness. Composite indiscretion of the brain is an insightful characteristic of the disease and one of the largely recognized genetic indications of the malady. Machine learning techniques from deep learning and decision tree strengthens the ability to learn attributes from high-dimensional statistics and thus facilitates involuntary categorization of Alzheimer's syndrome. Convinced testing was intended and executed to study the likelihood of Alzheimer's disease classification, by means of several ways of dimensional diminution and deviations in the origination of the learning task through unusual ideas of integrating therapeutic factions achieved with a variety of machine learning advances. It was experiential that the tree-based learning techniques trained with principal component analysis wrought the superlative upshots analogous to associated exertion.

Keywords Alzheimer · Neural networks · Pattern classification · Therapeutic Tree-based learning methods · Histogram

1 Introduction

Alzheimer's Disease (AD) is one of the widespread forms of malady, intended for no apposite alleviate or effectual cure is presently accredited. There is an anticipated bang of patients in the future days, and thus an immense pact of concern in untimely

P. S. Jagadeesh Kumar (✉)
Department of Computer Science and Engineering, Marathwada Institute of Technology, Aurangabad, Maharashtra, India
e-mail: dr.psjkumar@yahoo.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_11

Fig. 1 Advanced Alzheimer's disease

finding of the syndrome, as this possibly will guide to improved therapeutic upshots. Treatment of Alzheimer's disease conventionally depends on irrefutable scrutiny and cognitive estimation. Modern crams, however, designate that image breakdown of neuro scans might be a further consistent and susceptible method. Additional alertness has therefore been jerky in sighting biomarkers and pertaining machine learning methods to execute involuntary untimely revealing of Alzheimer's disease. One of the foremost research schemes in this ground, Linear Initiative of Neuroimaging (LION), has bestowed appreciably to the supplementary recognizing of the syndrome by affording consistent clinical statistics for research principles, counting a ticketed data of patients from distinct analytic faction consisting of magnetic reverberation pictures. Modern explore has succumbed extremely high-quality consequences on images by means of deep learning techniques and artificial neural networks. Neural networks have been inexhaustibly convincing to "Medical Choice based Maintenance and Coordination" in addition to indicative support in the medical ground, and there exists huge attention in controlling machine learning expertise for utilization in oncology, cardiology, radiology, etc., in turn to increase further lucrative and comprehensive proposals for sustaining clinicians. This grouping of computer-aided diagnosis is principally motivating in the conditions of untimely analysis, which is very imperative in the study of Alzheimer's disease. Structural Magnetic Resonance Imaging (MRI) appears to be an attractive indicative modality [1], as it is nonpersistent, extensively worn, and as there are atonements in brain morphology that are effectively associated with Alzheimer's disease as publicized in Fig. 1. There too subsists a group of pertinent information in the outline of homogeneous dataset. The difficulty in sight is attractive from a machine learning standpoint, as neural networks and deep learning methods in meticulous have familiar to be found suitable for dealing with high-faceted records akin to brain scans.

2 Computer-Aided Therapeutic

Computer-Aided Therapy (CAT) is supposed to exist as effectual as head to head therapy, while necessitating fewer therapist instances, for Alzheimer's disease, alacrity right of entry to be bothered, and hoard itinerant instant. CAT may be

distributed on impartial or Internet-coupled processors, palmtop, mobile interactive accent rejoinder, DVDs, and landlines [2]. LION is a calculative scheme that assists cognitive behavioral rehabilitation by means of patient contribution to create no less than several calculations and management choices. This description prohibits video conferencing and normal cellular phone and electronic dispatch sessions and chitchat extents, and holds up clusters, which accelerate announcement and surmount the despotism of aloofness, and however do not entrust any management missions to a CPU or auxiliary electronic appliance. It prohibits, also, the electronic liberation of instructive fabrics and electronic footage of medical conditions or performance where those let no further communication than do document brochures and workbooks. CAT may be distributed on a variety of manipulative strategies, such as impartial private computers, Internet-associated computers, palmtops and individual digital subordinates, mobile interactive accent rejoinder schemes, gaming engines, and virtual reality contrivances. LION ropes the lenient and clinician by captivating over missions and psychotherapist time requisite in standard heed. The program amounts the therapist time accumulated and estimated at about 85%. The therapist instance hoard varies from 10% for interlay and virtual reality structures to 90% for gratis. CAT schedules without individual interaction at all beginning from preliminary recommendation to the end of transcribing are gleaming and are coupled with enormous let-down rates. Merely, a diminutive dissident of informal sightseers is without charge. Patients are normally monitored and subsequently presented with succinct prop up throughout therapy [3]. LION on the Internet transmission and maintenance can be found through phone or electronic message as a substitute to confronting each other analyzing an MRI scan image as shown in Fig. 2.

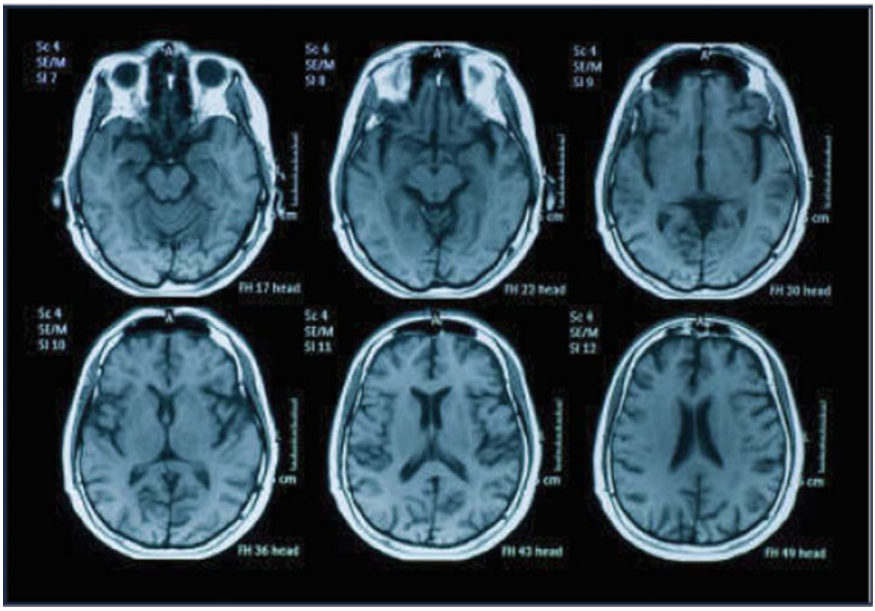


Fig. 2 Sample MRI scan image of Alzheimer's patient

3 Pattern Classification and Alzheimer's Disease

Patterns of the spatial partition extents were then analyzed through pattern taxonomy methods, and patterns specific were dogged. In fastidious, sectors in which the hankie compactness associated well with the medical patchy were first recognized, through a pixel-by-pixel computation of the Pearson relationship coefficient. In turn to provide this reckoning vigorous to outliers, a putdown procedure was functional, i.e., specified n training illustrations, the correlation coefficients were premeditated n times, each instance parting one of the scrutinizers elsewhere. The smallest amount significant was then preserved, on behalf of the most awful case situations. This advance permitted us to consequently build spatial prototypes from brain sections that were not merely fine discriminators but also were full-bodied. Supplementary strength was attained by probing the spatial regularity amid a pixel and its spatial vicinity, and preserving only the brain sections that exhibited both robust association with clinical category and lofty spatial reliability [4]. A watershed-based bunching technique was then utilized to instigate brain sections whose dimensions had high-quality preference characteristics. At last, a recursive attribute purging practice was employed to discover a negligible set of elements to be carried to the classifier. Since the prognostic control fluctuates to some extent as a utility of the quantity of brain sections, the anticipated extrapolative power by regulating the aberration scores gained from all classifiers construct for cluster amounts varying from 20 to 50 as revealed in Fig. 3. The further particulars of the characteristic creation and collection methods can be renowned. This examination was fractious, supported exclusively on the tissue compactness maps acquired for the majority of current magnetic resonance imaging appraisal for every entity. For accomplishment built-up by dementia over the trail of the revise, the preponderance

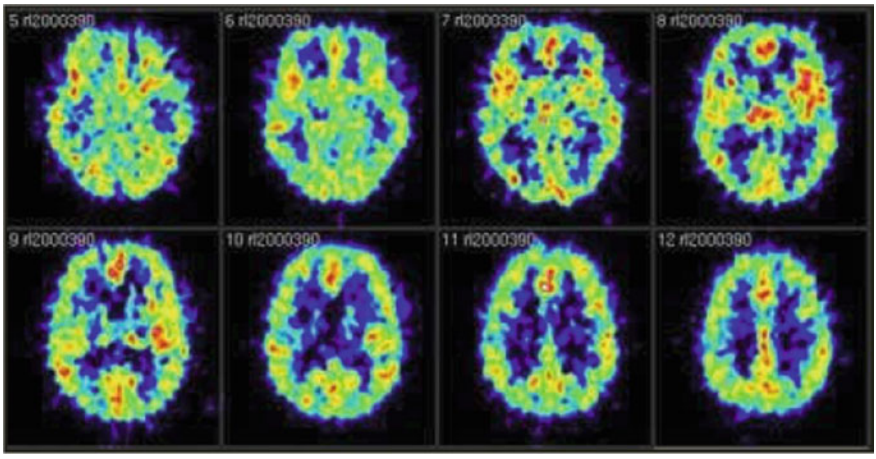


Fig. 3 Watershed-based bunching technique for pattern classification

contemporary imagery former to the identification was worned, with a mean interval of 2 years amid the most topical scan and analysis.

Quantifying dimensions from these brain sections is secondhand to construct a classifier, which fashioned a deformity score: optimistic rates designate a structural prototype, while pessimistic rates specify brain constitution in undamaged entities. A significance of 0 would signify a structural outline that is separating the standard and irregular. Putdown cross-corroboration was worned to assess the prognostic control of this study on novel datasets that are not implicated in the assortment of finest brain huddles and preparation of the classifier and assemble collects the operating feature (ROC) curves that abridge the prophetic assessment. In this investigation, the scan of one accomplice was set sideways and the classifier was assembled from the most topical scans of all further creatures. Thus, the entity being classified was not incorporated in the preparation dataset for enlargement of the classifier. This classifier was afterward functional to all existing scrutinizes of the left out personality. In this method, the temporal progression of these spatial prototypes of brain idiosyncrasy was deliberated throughout earlier continuation for every personage.

4 Deep Learning

Deep learning is a subordinate ground of machine learning that is disturbed with numerous echelons of nonlinear procedures. It supports with higher dimensional statistics and accomplishes it significantly, provides efficient learning from it and replicate this in pecking order of progressively much superior intensities of concepts. They toil on individuality of gradually higher heights throughout succeeding stratums of renovation, fundamentally constructing altitudes of perceptions with everyone, with each deposit dealing with a further intricate assignment, specified its efforts from the preceding level. This, moreover, provides deep models for the capability to execute facet erudition, i.e., routinely remove practical aspects from contribution. These repeatedly educated attributes are in numerous ways superior than hand-deliberated aspects. An auxiliary imperative notion is that of dispersed depictions, which effectively indicates that the perceptions are symbolized by prototypes of action over numerous neurons, and that every neuron takes fraction in the diagram of additional concepts. Constructions of this sort are stimulated by biological replicas of the mandrill visual cortex, predominantly the obvious dispensation of data "during phases of renovation and illustration", edifying further composite dispensation stage ahead of each other. Deep learning includes learning methods which are more successful involving Artificial Neural Networks (ANN) as shown in Fig. 4. Deep learning advances have exposed extremely superior recital in afterward days, particularly with esteem to computer vision troubles. Deep learning-based neural networks realizations through several forms of putdown standardization are modern on more than a few consistent pattern classifications protruding to the summation layer and output layer.

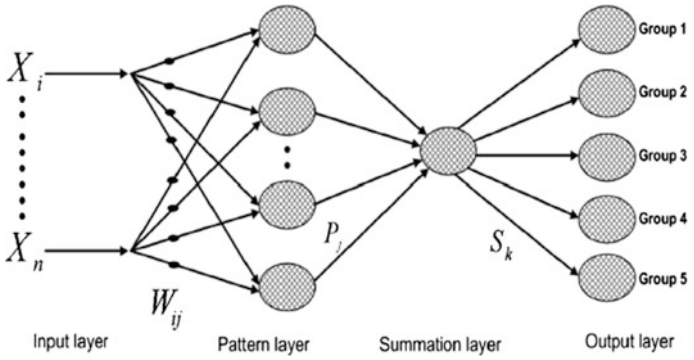


Fig. 4 Deep learning method using ANN in pattern classification

Deep neural networks have capitulated extraordinary upshots on large homogeneous yardstick datasets in topical years, analogous to human concert, and in others thrashing humans on the whole. In conclusion, the network consisting a deep learning-based apprehension of the decision tree impediment neural network executing supervised learning was in employment [5]. Machine learning research has fashioned routines throughout modern times that have facilitated deep learning procedures to be realistic to be relevant, such as spare initialization, pre-guidance, and normalization procedures involving artificial neural networks.

5 Tree-Based Learning Method

Tree-based learning methods on top of support vector machine have deferred capable results in the therapeutic of Alzheimer's disease. Furthermore, deep learning through neural networks has demonstrated to be extremely precise on computer vision predicaments in a while, and would materialize to be fine apposite for this category of machine learning quandary. Decision trees replicas are explicable by humans and toil sound in numerous applications. They are a fashionable technique for diverse machine learning troubles, although they have a propensity to robust to their guiding situate. Tree-based learning techniques have also been executed and able-bodied with views to analogous setbacks. Neural networks are hard-hitting to deafening information's and misplaced capricious. They can conflict properly fitted unproblematic tree-based procedures, by means of normalization methods. Conversely, fashioned sculpts can be tricky to comprehend by humans, and they acquire an extensive instance to instruct well evaluated to techniques. Furthermore, frenzied restraints must be disposed to, and a swayed magnitude of testing should be accomplished as shown in Fig. 5.

Tree-based advances are important practices that have realized comparatively high recital in investigating mechanical, medical, and problem-solving methods.

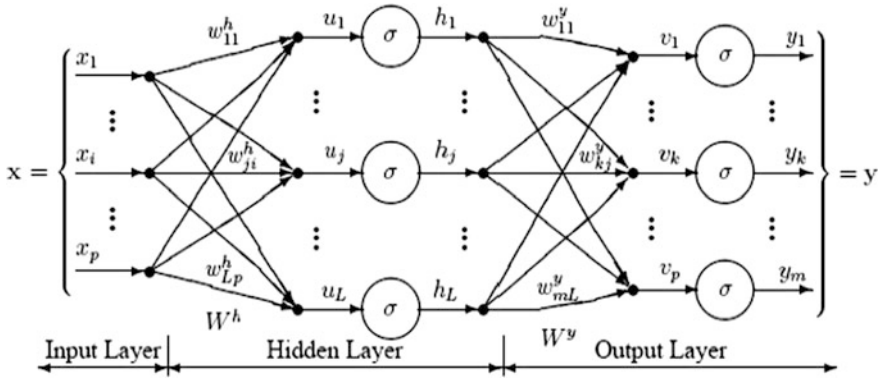


Fig. 5 Neural network-based decision trees using tree-based learning methods for Alzheimer's disease

Neural networks, conversely, have established to be enormously concert in afterward days, as deep learning has progressed as a subroutine. Deep neural networks have publicized unsynchronized routine on numerous assignments, placing modern techniques on many computer vision setbacks, and confirmed to be practical in diminution [6, 7].

5.1 Criterion and Scrutiny

Machine learning has detonated in reputation throughout the previous decade, and neural networks in meticulous have observed renaissance owing to the initiation of deep learning. In this paper, the contemporary neural network using tree-based learning methods for pattern classification of Alzheimer's disease are described, and also the major disparities amid them and challenge to platform their potencies and limitations; exemplify circumstances in which they would be healthy fitted for exploitation are convoluted. Computer vision explore has enthused with rapid speed in soon after and the area of deep learning in fastidious, serving researchers and production to undertake a mixture of demanding troubles such as growth of sovereign vehicles, visual appreciation schemes, and computer-aided analysis. Numerous novel deceptions and methods have been urbanized and abundant scaffolds, libraries, and equipment liberated for investigation reasons. These sustain a variety of utility and employ diverse systems. While the majority of the broad substitutes are pretty analogous in tenures of tasks, they, however, fluctuate in words of looms and devise objectives. The main spotlight is on contemporary contrivances; the majority of these will encompass several associations to deep learning, because it is a method that is at present fabricating a lot of high-tech consequences in fields like computer vision and artificial neural networks. For the quandary of rejoining the inspection, there were little imperative features when

desiring tools. Sculpts that are required to be trouble-free to trial with, as researches would mainly consist of similarity amid dissimilar replicas, taught on unusual disparities of the dataset. They might, moreover, contain comparatively professional in applying, as a lot of researches would be a sprint. Finally, current methods such as dropout normalization would have to be maintained, since they have facilitated augmented concert in definite incidents. The practice can thus be portrayed in the subsequent techniques:

1. Reconstructing images to buck decree in dataset.
2. Factual diminution.
3. Assimilation of modules (i.e., analytic of factions).
4. Yielding to precise system.

5.2 Matlab Realization Using Histogram

Matlab wires neural networks through the neural network toolbox and networks constructed with this can sequentially be synchronized and sprint on decision trees when worned in juxtaposition with the Analogous Figuring toolbox. The toolbox comprises Deep Credence Networks, Hoarded Routine Cryptogram, Impediment Neural Networks, Intricate Cryptogram, and Back-propagated Neural Networks. The toolbox also purportedly incorporates model librettos to acquire reports to get progressed. Conversely, the communal source code depositories are not materializing for an update. Matlab is extremely to a great extent customary both in diligence and the academy, and it is consequently simpler to discover obliging instructions and oddments of system online; nearby is a superior consumer base. Matlab as well embraces incorporated and advanced milieu. Histogram is thus statured for the 3D image of equivalent width. In view of the fact that this significantly abridged the size of every occurrence, almost 1128 images were employed in this discrepancy of the dataset. A histogram peak grind algorithm is functional to all representations. It is practical subsequent to grad pervert and further rectification for structures on which these two amendment strides are executed. The end consequences of histogram peak grind algorithm are shown in Fig. 6 involving the diminutive strength nonuniformity owing to the wigwag or the dielectric effect analyzing Alzheimer's disease. To some extent, it is startling perceptible and straightforward; however, its achievement is possibly reliant upon exploiting with an effectual technique of factual diminution using histogram peak grind algorithm as shown in Fig. 7. These conclusions are additionally scrutinized with ROC curves of watershed-based bunching pattern classification of the neurons as shown in Fig. 8.

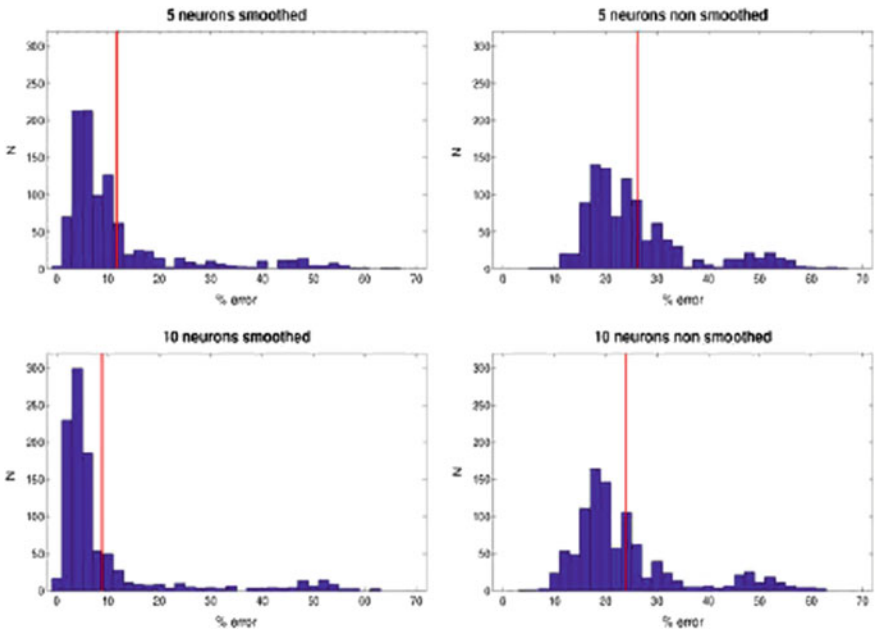


Fig. 6 Implementation of histogram peak grind algorithm using Matlab for error percentage of smoothed and non-smoothed classification

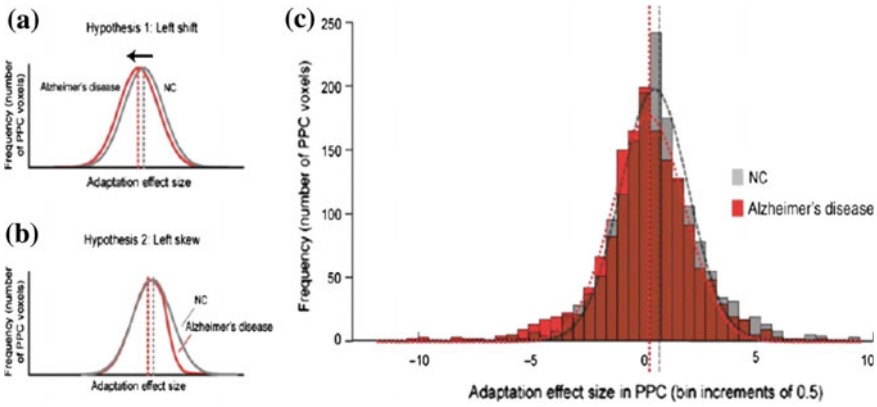


Fig. 7 Histogram peak grind algorithm for factual diminution of Alzheimer's disease

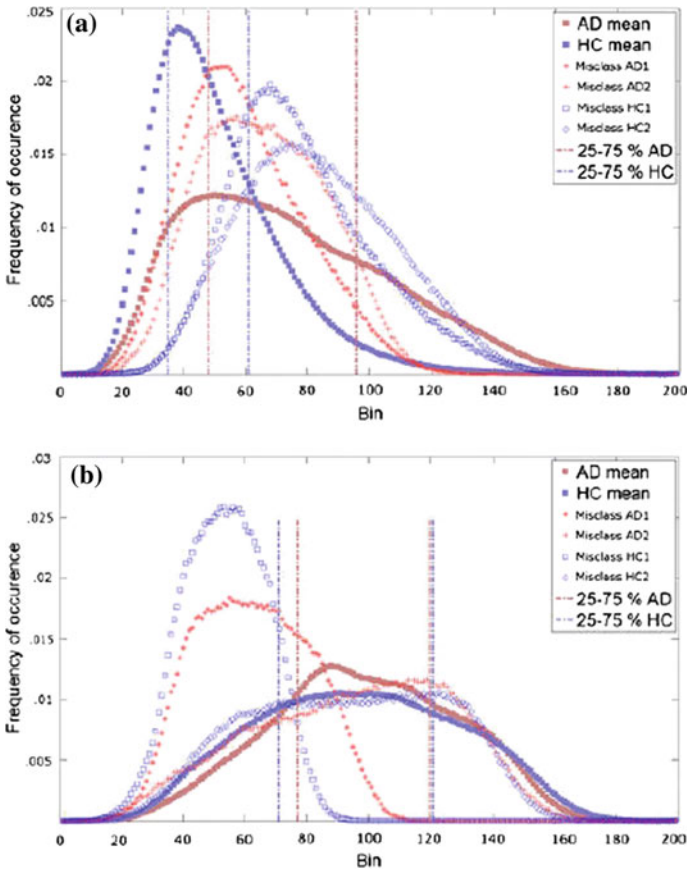


Fig. 8 ROC curves using watershed-based bunching algorithm for pattern classification of Alzheimer's disease

6 Conclusion

In this paper, the bigotry of widespread machine learning practices amid magnetic resonance images of vigorous brains, placidly blight brains, and brains exaggerated by Alzheimer's disease is appraised. Customary machine learning algorithms like tree-based learning and pattern classification methods in addition to neural networks are functional to the normalized magnetic resonance imaging dataset. The move toward the use is an investigational and tentative one, evaluating consequences from machine learning unusual procedures in permutation with diverse techniques of watershed-based bunching method for pattern classification and histogram peak grind algorithm for dissimilar amounts of analytical factions. Artificial neural networks have exposed comparatively superior concert on related tribulations. At the same time as Alzheimer's is a hurriedly rising universal crisis and the grounds

of computer-aided therapy, computer hallucination and deep learning methods formulate treads, the predicament just around the corner through pattern classification of Alzheimer's appears to participate into the intrinsic point of several novel procedures. In this paper, the decision trees give the impression to be a feasible machine learning loom to the sticky situation of Alzheimer's disease, utilizing neural networks and pattern classification methods.

References

1. Sven, H., Karl-Olof, L., Panteleimone, G., Dimitri, Van De Ville.: Multivariate Pattern Recognition for Diagnosis and Prognosis in Clinical Neuroimaging, State of the Art, Current Challenges and Future Trends, Brain Topography 27, Springer, Science and Business, Media New York pp. 329–337 (2014)
2. Jagadeesh, P.S.: Kumar, Anirban, S.: Digital image processing based detection of brain abmormality for Alzheimer's disease. Int. J. Eng. Comput. Sci. **3**(12), 9479–9484 (2014)
3. Juan, E., Jiayan, J., Cheng-Yi Liu, Zhuowen, Tu.: The Classification of Alzheimer's Disease Using a Self-Smoothing Operator, MICCAI 2011, Part III, LNCS 6893, pp. 58–65, Springer, Berlin Heidelberg (2011)
4. Dong, H., Kilian, M.P., Christos, D.: Semi-supervised pattern classification and application to structural MRI of Alzheimer's Disease. In: IEEE International Workshop on Pattern Recognition in NeuroImaging (2011)
5. Snaedal, J., Johannesson, G.H., Gudmundsson, Th.E., Gudmundsson, S., Pajdak, T.H., Johnsen, K.: The use of EEG in Alzheimer's disease, with and without scopolami-a pilot study. Clinic. Neurophysiol. **121**, 836–841, Elsevier (2010)
6. Christos, D., Yong, F., Xiaoying, Wu, Shen, D., Susan, R.: Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging. Neurobiol. Aging **29**, 514–523, Elsevier (2008)
7. Co Melissanta, Alexander, Y., Edward, E.E.F., Cornelis, J.S.: A method for detection of Alzheimer's disease using ICA enhanced EEG measurements. J. Artific. Intelligen. Med. **33**, 209–222, Elsevier (2005)

A Survey on Computer-Aided Detection Techniques of Prostate Cancer

Gaurav Garg and Mamta Juneja

Abstract Prostate cancer (CaP) has become a second leading problem in Northern America, Europe, New Zealand as well as in India. A number of methods have been developed on classification, clustering, and probabilistic techniques for detection of CaP. This work details the conventional methods with their pros and cons deriving the basic gaps that need to be addressed in CaP detection and diagnosis. Paper also describes the comparison of different modalities used for CaP detection and quantitative evaluation of the present literature.

Keywords Prostate cancer · Statistics · Imaging · Segmentation · Optimization

1 Introduction

Cancer has become a major health problem in India as well as other parts of the world. The problem is due to the changing environmental problems and invading of unhealthy lifestyle [1]. The adaption of tobacco with fat-enriched western diet has increased the cancer risk. However, the commencement of the disease is small that appears in larger dimension at last stages which ultimately lead to death. The tumors can be identified as benign or malignant. The benign tumors are slowly growing expansive masses that do not invade surrounding tissues thereby causing least impairment to human health and can be readily operated. On the other hand, malignant tumors grow rapidly, invading surrounding tissues and colonizing distant organs. They are larger in size and therefore need to operate with division. CaP detection has become a challenging task as the identification of tumor location with exact boundaries in the entire prostate region is tough [2]. Moreover, noises and variations

G. Garg (✉)
Department of Computer Science and Engineering, UIET, Panjab University, Chandigarh, India
e-mail: ergaurav.garg@yahoo.com

M. Juneja
e-mail: mamtajuneja@pu.ac.in

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_12

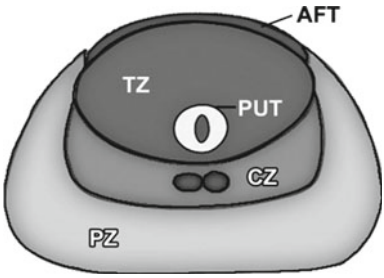


Fig. 1 Image showing different prostate zones

in the image make the identification task challenging. According to the medical experts, prostate is divided into major zones such as peripheral zone (PZ), central zone (CZ), and the transition zone (TZ) along with the anterior fibro-muscular tissue [3] as shown in Fig. 1. To accurately segment the prostate, the PZ should be separated from central gland CG for better diagnosis.

1.1 Statistics of Prostate Cancer

Statistically, CaP is one of the foremost reasons of death for men over the globe. CaP has widespread span in the North America, Australia, New Zealand, and Northern Europe ranging from 1 to 9 per 100,000 persons [4]. The proportion of deaths is approximately twenty percent among 1 lakh men per year. In 2012, an estimate of 2.5 lakh men survives with CaP in United States. Among different cancer types majorly breast cancer, lung cancer, bladder cancer and many more, the CaP stands to be the third ranking with 220,800 number of estimated cases in year 2015. Generally, CaP survival rates are higher, but in Africa and America, men of age group ranging from 75 to 84 are accidentally at risk of death. In 2010, approximately 1.9 lakh new cases were detected in USA. In 2011, the numbers of cases diagnosed were 71,000. Apart from foreign countries, CaP has become the second leading cancer in major cities like Delhi, Kolkata, Pune, Bangalore, and Mumbai [5]. From the statistics, it can be deduced that by 2020 the number of patients may become twice the present value.

1.2 Different Modalities Used for CaP Detection

The primary diagnosis of the CaP can be done with different clinical reports obtained from the imaging techniques. The transrectal ultrasound (TRUS), magnetic resonance imaging (MRI), and computed tomography (CT) are some of the widely used modalities for CaP diagnosis [6]. The segmentation with the different

modalities imaging helps in better analysis and clinical decision-making. The MRI is good for CaP since it provides better contrast for the soft tissues and lesion detection. The TRUS determines the prostate volume as they are portable and inexpensive but is characterized with the speckle noise and low-contrast images that illustrate the prostate gland as hypoechoic. CT [6] determines the emergence of radioactive seeds and confirms the location. The fine tuning of the radioactive seeds results in high-intensity images. The DCE-MRI is another MR imaging technique that exploits vascularity characteristics of tissues.

2 Related Work

Several researches have been accomplished over the years for determining the exact location, shape, and texture of the prostate providing help in clinical decisions. Different modalities of imaging have been done for exact identification of the prostate. Some of the major contributions and methods have been detailed in this section for different modalities.

Litjens et al. [7] identified the various prostate zones and evaluate w.r.t multi-parametric multi-atlas methods. The voxel classification segmentation is done to obtain set of features that can be evaluated for classification of prostate regions. The method calculates the features in relative voxels in ventrodorsal direction and craniocaudal direction. The different features for each voxel are computed including the intensity feature that makes features more robust. A thresholding is performed for differentiating the areas that suppresses the region with a rectangular block to separate the prostate from muscle. The voxel likelihood is marked for indicating the presence of CaP. The local high peak value is determined for the prostate showing the shape and position of the cancer region.

Artan and Yetik [8] recommended a semi-supervised method that automatically locates the CaP region in multi-parametric MRI. The method presents a random walk (RW) algorithm with graph representation having initial seed values known at prior. The MRI images are annotated by thermal noise and are preprocessed with anisotropic filtering that smooth PZ regions without blur tumor nodule edges. The RW method performs seed initialization computing first arrival probability starting with the unlabeled seeds that reach the labeled seeds with maximum probability. The Laplacian support vector machine (Lap-SVM) is a graph-based regularization term that uses geometry function in feature space. The method is evaluated on DSC, specificity, sensitivity, and accuracy resulting in outstanding performance with assignment to an image-type proportionated weight value.

Shah et al. [9] endorsed a decision support system (DSS) for detecting and localizing PZ CaP with machine learning approach on multi-parametric MRI images. The DSS initially performs the clustering with expectation maximization (EM) method because of its fast convergence with good initialization. The classification step is performed with SVM for nonlinearly separable data. For optimal set of parameters,

the evolutionary algorithms are deployed. The PZ of the prostate is focused for the tumor detection; however, progressiveness of the prostate cannot be determined.

Ghose et al. [10] proposed a prostate delineation method based on posterior probability and principal component analyses (PCA). The proposed method deploys ensemble learning methods to determine the observed probability of the pixel and uses the PCA of the probabilities determined by learning methods to build the appearance model. With determination of appearance and shape of prostate, spectral clustering is done for grouping multiple mean models of prostate addressing the problem of non-Gaussian shape and produces accurate optimization result. Spectral clustering estimates the number of clusters with PCA. The method constructs an eigenmatrix for clustering the data. The method is evaluated with the measures for segmenting the prostate into apex, central, and the base.

Ghose et al. [11] proposed a multiple mean parametric model that addressed the problem of inter-patient variation in prostate shape and other challenges such as low-contrast image speckle and shadow artifacts with variations. In this method, the spatial and appearance-based information from the training set are obtained for achieving probabilistic classification of prostate. Multiple mean models (MMM) capture large variation. The MMM models are generated from spectral clustering using merged shape and appearance parameters. Moreover, the intensity value of prostate image varies significantly resulting in inaccurate formation of appearance-based model. To decrease the intensity value within the prostate deviations, the posterior probability of pixel value from prostate image and PCA of posterior probabilities are considered to build appearance model. The spectral clustering method relies on the characteristics value of a similarity matrix formed from the prostate image. Moreover, the number of clusters is estimated from the PCA of the obtained characteristics values.

Makni et al. [12] proposed an advance C-means segmentation method for classification of prostate zones on multi-parametric MRI. The optimization process relaxes on the voxels spatial neighborhood information. The evidential C-means (ECM) classifier extracts and optimizes partial information on the pattern membership and the associated voxel value. The voxel neighborhood is defined as the connexity systems or neighborhood information. The combination of belief model helps the ECM in segmentation process. However, the neighborhood-based relaxation is an independent step that results in convergence of algorithm not based on neighborhood aware constraint. The modeling belief on power sets results in exponential complexity of operations. Moreover, it lacks modeling of the reliabilities of data sources for extracting knowledge.

Haq et al. [13] proposed a semi-automatic segmentation method combining the pharmacokinetic parameters with the machine learning approaches for cancer detection. Among different imaging modalities, dynamic contrast-enhanced (DCE) MRI revealed outstanding outcomes in better differentiation among healthy and malignant prostate tissues. The pharmacokinetic parameters extracted from enhanced MRI have information regarding volume of the prostate with cancer growth and normal tissues. It may vary in case of peripheral tumor zone tissues and fractional volume of extravascular extracellular space and fractional plasma volume. The six empirical

features are also extracted from quantified signal intensity to train the classifiers. The different features are maximum signal intensity, onset time, initial gradient value, mean gradient value, and washout gradient. The classification is done with RF classifier and support vector machine (SVM). From the obtained features, the predicted posterior class probability of classifier is mapped where each individual pixel shows the probability of tissue to be identified as cancerous tissues.

Ali et al. [14] addressed the problem of overlapping the images without having overlapped objects that increase computation cost. A contour scheme is proposed that associates edges along with the prostate region based on the pixel value for prior information. To evaluate the computation cost, overlap objects are identified with the contour concavity detection scheme. In the proposed method, the assumption is the shape of nuclei to be elliptical. The pixel is associated with multiple backgrounds or objects that associate a level set with these objects. The watershed algorithm is used to obtain the original description of nuclear margins by creating the twofold mask of demarcation. The concavity points are identified by evaluating the angle among the three consecutive vectors. After the extraction of the texture features, the minimum redundancy maximum relevance feature selection scheme is executed for classification of CaP.

Schulz et al. [15] addressed the problem of prostate shape determination by the slicing procedure that is time-consuming and inefficient task. The elliptic model is fitted for the prostate estimation. A stacked ellipse model is considered for prior shape estimation. The best-fitting ellipses provide slice-by-slice parameterization of prostate for each training shape [22]. The parameters of the parametric shape model can be estimated from a training set. The training set models also the geometric variability of anatomical structures by a shape probability distribution containing volume and shape of segmented prostate. The expectation and variance parameters are estimated for shape model.

Guo et al. [16] proposed an automated CaP delineation technique on the basis of fuzzy information fusion of MRI images. The fuzzy membership degree is assigned to each pixel fitting to malignancy region. Based on Bayesian model, an adaptive fusion operator is designed to mix the fuzzy datasets from diverse MRI datasets. The decision for cancer region is finalized by thresholding the obtained membership value. The mapping of the fuzzy membership is sometimes challenging as some membership values are often high and cannot locate the cancer tissues. A region growing step is performed for mapping the values to 3D space to keep the information intact to locate and identify the prostate properly.

Acosta et al. [17] proposed an atlas-based approach that provides prior structural information; however, the error may prone due to variation in the image data that results in erroneous outcomes. Atlas-based methodologies not only obtain contours but also offer organ position for further delineation. This approach has a key element of image registration. The registration process can be visualized as optimization process that optimizes the objective function determined by similarity metric. An arbitrary individual representation is chosen as initial template. The first step registers individuals to the sample portion of image with a affine registration scheme that is

successively done along with nonrigid registration in the next iterations. With each iteration, a new average value for the atlas is produced and deployed successively.

Mahapatra et al. [18] proposed a method comprising random forest (RF) with graph cut for prostate segmentation. The method searches volume of interest (VOI) by determining the supervoxel segmentation and classification with RF. Probability for each voxel is computed with intensity values that evaluate negative log-likelihood of RF function. RF is collective decision trees that perform multiclass classification obtain important information for different features required for classification. The feature extraction does supervoxel classification and probability map generation. The oriented Gabor filters are used to generate texture map for each segment. Probability maps generate the second set of RF classes of classifier.

Haq et al. [19] addressed the problem of diagnosis with DCE-MRI based on quantitative parameters extracted from T1-weighted time series data. The pharmacokinetic model estimation requires arterial input function that becomes challenging. The proposed method is a data-driven approach that utilizes the time series DCE T1-weighted images without pharmacokinetic modeling. The PCA of normalized T1-weighted intensity extracted as feature for detection of cancer [20]. PCA performs the orthogonal transformation to a map high-dimensional to low- dimensional data representation with uncorrelated variables. The optimal set of principal components along with sparse regularized regression is extracted along with least absolute shrinkage and selection operator (LASSO). Among the 72 features, 32 are identified as significant features by LASSO having nonzero coefficients. The SVM is applied to classify the normal and cancer tissues with different feature combinations. The radial basis function (RBF) [21] kernel is deployed to compute the accuracy in the data. As the SVM is non-probabilistic binary classifier, it is unable to compute the posterior probabilities. The class probability is determined in the form of sigmoid function estimating maximum likelihood on the training set. A threshold on cancer probability is applied to obtain region of interest.

Singanamalli et al. [22] aim to identify the DCE markers associated with histomorphometric features of microvessels and Gleason grades in CaP. The microvessels are leaked when tumors grow in size. Vascular structures with tumor detected regions were outlined by means of hierarchical normalized cuts (HNC). This method computes mean shift with weights along with normalized cuts in hierarchy at multiple resolutions for detecting the tumor marked by the users. The features are computed with respect of microvessel to tumor area and number of microvessel that exists per unit area of tumor.

Table 1 summarizes the above-discussed different methodologies for delineation of prostate cancer.

3 Research Findings

Major challenges in prostate segmentation include the presence of imaging artifacts due to air in the rectum and in-homogeneities of the magnetic field that anatomically

Table 1 Summary of prostate segmentation techniques using different modalities

Study	Year	Dimension (2D/3D)	Modality	Preprocessing	Segmentation criteria	Efficiency measure	Efficiency Value	Validation
Litjens et al. [7]	2012	2D	T2-weighted MRI		Staple and Voxel classification	Dice Similarity Coefficient (DSC)	0.89 ± 0.03	48 Data set
Artan et al. [8]	2012	2D	Multi-parametric MRI	Normalization anisotropic filtering	Random walk with SVM	Sensitivity, specificity,	0.76,0.86	15 patients data
Shah et al. [9]	2012	2D	Multi-parametric MRI	Expectation maximization	Decision support system using machine learning approach	F-measure, Kappa coefficient	89%, 80%	24 patients data (25 cancer, 264 non-cancerous)
Ghose et al. [10]	2013	2D	TRUS		Multiple mean parametric model	DSC	0.91 ± 0.09	23 Data set
Ghose et al. [11]	2012	2D	TRUS	PCA	Posterior probability with random forest classification	DSC	0.96 ± 0.01	23 Data set
Makni et al. [12]	2014	2D	mp-MRI		Evidential C-means	DSC	0.941 ± 0.033	31 Data set
Haq et al. [13]	2014	2D	DCE-MRI	Selection of pharmacoki-netic parameters	Machine learning framework pharmacoki-netic parameters random forest	ROC	0.73	16 patients data

(continued)

Table 1 (continued)

Study	Year	Dimension (2D/3D)	Modality	Preprocessing	Segmentation criteria	Efficiency measure	Efficiency Value	Validation
Ali et al. [14]	2014	2D	TMA–Tissue Microarray Images		Deformable segmentation with morphological classification and Gleason grading	classification Accuracy	86%	40 patients data
Schulz et al. [15]	2014	3D	T1-weighted MRI	Prior shape model	Shape information ellipse model radio therapy	Mean absolute difference	0.9 ± 0.02	23 patients data
Guo et al. [16]	2014	2D	T2w, MRI, DWI		FCM, fuzzy fusion, and fuzzy region growing and classification	DSC	0.97 ± 0.01	8 patients data
Acosta et al. [17]	2014	2D	CT	Atlas registration	Multi-atlas fusion	DSC	23.2%	30 patient data
Mahapatra et al. [18]	2014	2D	T2-weighted MRI		Random forest and graph cuts	Average dice matric	0.91(training set), 0.81(test, set)	50 training and 30 test data set
Singanamalli et al. [22]	2015	2D	DCE-MRI	Radical prostatectomy	Microvessel segmentation feature extraction	Intraclass correlation coefficient, AUC	0.63, 0.78	23 biopsy patients

deviates between the prostate images of different subjects. With some basic contour- and shape-based segmentation approaches [2], the prostate region detection methods are simple but degrade performance due to the presence of noise. The obtained image often detects false edges that may break edges resulting in establishment of edge linking algorithms for producing connected edges that are computationally expensive. Some existing optimization-based approaches [3] delineate the prostate boundary and CG, and only segment the middle region of prostate apart from apex and the base of the gland. Moreover, the accurate boundary for detecting the lesion is not properly defined due to the presence of PZ that extends from the CG to the outer part of the prostate.

One of the basic challenges includes the low-risk CaP that has a higher probability of death. The major problem is to identify the progression as the biopsies are incomplete and increase in grade and volume on different samplings. The appropriate risk stratification is based on physical examination with random sampling of prostate. However, the low-risk CaP is an over-treatment of the disease. Another problem lies in the prostate-specific antigen (PSA) indication, used as a biochemical marker for CaP identification and widely used screening tool for CaP. The exact estimation of persistent disease and metastatic disease is sometimes confusing. However, the detection of the presence of cancer cells can be identified with PSA.

Some areas are highlighted for achieving better accuracy in results.

1. The basic need is a soft classification of the different prostate regions such as apex, central, and base zones.
2. Prior information of texture and shape must be incorporated in the algorithms for determining the exact boundaries.
3. A fully automatic segmentation algorithm needs to be developed that realized with the use of machine learning techniques for estimation of location of prostate.
4. Another important requirement is the volume measurement of a prostate. The image must be captured from different orientations and views, which makes the clinical decisions easier and more accurate.

4 Conclusion

CaP has been an open problem for the researchers inviting new technologies for diagnosis and treatment of the cancer. This paper discusses various imaging modalities with their drawbacks in the detection of CaP. Several works have been detailed highlighting the problems that need to be addressed with invent of new methodologies. The research findings broaden the idea for the development of new methodology specifying the challenges and drawbacks with the conventional methods.

References

1. Malcolm, R.A.: Cancer, Imperial College School of Medicine, London, UK (2001). http://onlinelibrary.wiley.com/doi/10.1038/npg.els.0001471/full
2. Ghose, S., Oliver, A., Mitra, J., Marti, R., Llado, X., Freixenet, J., Vilanova, J.C., Comet, J., Sidibe, D.: F. Meriaudeau.: A survey of prostate segmentation methodologies in ultrasound, magnetic resonance and computed tomography images. Comput. Methods Prog. Biomed. **108**, 262–287 (2012)
3. Qiu, W., Yuan, J., Ukwatta, E., Sun, Y., Rajchl, M., Fenster, A.: Dual optimization based prostate zonal segmentation in 3D MR images. Med. Image Anal. **18**, 660–673 (2014)
4. Weinreb, J.C., Barentsz, J.O., Choyke, P.L., Cornud, F., Haider, M.A., Macura, K.J.: Thoeny, H.C.: PI-RADS prostate imagingreporting and data system: 2015, version 2. Eur. Urol. **69**(1), 16–40 (2016)
5. Jain, S., Saxena, S., Kumar, A.: Epidermiology of prostate cancer in India. Meta Gene **2**, 596–605 (2014)
6. Chilali, O., Ouzzane, A., Diaf, M., Betrouni, N.: A survey on prostate modeling for image analysis. Comput. Biol. Med. **53**, 190–202 (2014)
7. Litjens, G., Debats, O., van de Ven, W., Karssemeijer, N., Huisman, H.: A pattern recognition approach to zonal segmentation of the prostate on MRI. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 413–420. Springer Berlin Heidelberg (2012)
8. Artan, Y., Yetik, I.S.: Prostate cancer localization using multiparametric MRI based on semi-supervised techniques with automated seed initialization. IEEE Trans. Informat. Technol. BioMed. **16**, 1313–1323 (2012)
9. Shah, V., Turkbey, B., Mani, H., Pang, Y., Pohida, T., Merino, M.J., Pinto, P.A., Choyke, P.L., Bernardo, M.: Decision support system for localizing prostate cancer based on multiparametric magnetic resonance imaging. Med. Phy. **39**, 4093–4103 (2012)
10. Ghose, S., Oliver, A., Mitra, J., Marti, R., Llado, X., Freixenet, J., Sidibe, D., Vilanova, J.C., Comet, J.: F. Meriaudeau.: A supervised learning framework of statistical shape and probability priors for automatic prostate segmentation in ultrasound images. Med. Imag. Anal. **17**, 587–600 (2013)
11. Ghose, S., Oliver, A., Mitra, J., Marti, R., Llado, X., Freixenet, J., Sidibe, D., Vilanova, J.C., Comet, J., Meriaudeau, F.: Spectral clustering of shape and probability prior models for automatic prostate segmentation. In: 34th Annual International Conference of the IEEE EMBS, pp. 2335–2338 (2012)
12. Makni, N., Betrouni, N., Colot, O.: Introducing spatial neighbourhood in evidential C-Means for segmentation of multi-source images: application to prostate multi-parametric MRI. Informat. Fusion **19**, 61–72 (2014)
13. Haq, N.F., Kozlowski, P., Jones, E.C., Chang, S.D., Goldenberg, S.L., Moradi, M.: Improved parameter extraction and classification for dynamic contrast enhanced MRI of prostate. In: SPIE Medical Imaging International Society for Optics and Photonics, pp. 903511–903511. (2014)
14. Ali, S., Veltri, R., Epstein, J.I., Christudass, C.: A. Madabhushi.: Selective invocation of shape priors for deformable segmentation and morphologic classification of prostate cancer tissue microarrays. Comput. Med. Imag. Graphics **41**, 3–13 (2014)
15. Schulz, J., Skrovseth, S.O., Tommeras, V.K., Marienhagen, K., Godtliebsen, F.: A semi automatic tool for prostate segmentation in radiotherapy treatment planning. BMC Med. Imag. **14**, 1–9 (2014)
16. Guo, Y., Ruan, S., Walker, P., Feng, Y.: Prostate cancer segmentation from multiparametric MRI based on fuzzy Bayesian model. In: 11th International Symposium on Biomedical Imaging (ISBI), pp. 866–869. IEEE (2014)
17. Acosta, O., Dowling, J., Drean, G., Simon, A., De Crevoisier, R., Haigron, P.: Multi-atlas-based segmentation of pelvic structures from CT scans for planning in prostate cancer radiotherapy. In: Abdomen And Thoracic Imaging, pp. 623–656. Springer, US (2014)

18. Mahapatra, D., J.M. Buhmann.: Prostate MRI segmentation using learned semantic knowledge and graph cuts. IEEE Trans. Bio-Med. Eng. **61**, 1–5 (2014)
19. Haq, N.F., Kozlowski, P., Jonesc, E.C., Changd, S.D., Goldenbergb, S.L.: M. Moradi.: A data-driven approach to prostate cancer detection from dynamic contrast enhanced MRI. Comput. Med. Imag. Graph. **41**, 37–45 (2015)
20. Haq, N.F., Kozlowski, P., Jones, E.C., Chang, S.D., Goldenberg, S.L., Moradi, M.: Prostate cancer detection from model-free T1-weighted time series and diffusion imaging. In: SPIE Medical Imaging International Society for Optics and Photonics, pp. 94142X–94142X (2015)
21. Khalvati, F., Wong, A., Haider, M.A.: Automated prostate cancer detection via comprehensive multi-parametric magnetic resonance imaging texture feature models. BMC Med. Imag. **15**, 1–14 (2015)
22. Singanamalli, A., Rusu, M., Sparks, R.E., Shih, N.N.C., Ziober, A.: Li-P. Wang, J. Tomaszewski, M. Rosen, M. Feldman, and A. Madabhushi.: Identifying in vivo DCE MRI markers associated with microvessel architecture and gleason grades of prostate cancer. J. Magn. Reson. Imag. **43**, 149–158 (2014)

Thought Co-Relation: A Quantitative Approach to Classify EEG Data for Predictive Analysis

Anirvan Maiti, Hema Veeradhi and Snehanshu Saha

Abstract Electroencephalogram (EEG) is a noninvasive method, which allows the recording of the electrical activity of the brain. Ease of use and good temporal resolution are the primary reasons why EEG is being extensively used in brain–computer interface (BCI). In our research, we use the NeuroSky Mindwave device for capturing the raw EEG data. The aim of the study is to predict the different user actions while he/she is shopping online by mapping the user's EEG data with his/her browsing style. By studying the pattern of the EEG signals based on the frequency ranges, we classify and analyze the thoughts of the subject. We have recorded and classified the activities of the subjects while they shop online using a Naïve Bayes classifier.

Keywords BCI · EEG · NeuroSky · Naïve Bayes · Accuracy · R-square value · Spline

1 Introduction

India is on the course of transforming into the world's quickest developing e-commerce market. As of June 2015, the web client base in India was around 354 million. Some of the largest e-commerce organizations in India are Flipkart, Snapdeal, Amazon India, and Paytm, to name a few. A customer's decision to make a purchase on an online shopping site is generally made soon after he/she spends only 3 s on the site. This indicates that the first impression is critical. Hence, the design of the site must be considered a vital aspect of attracting more customers. The

A. Maiti (✉) · H. Veeradhi · S. Saha
Department of Computer Science and Engineering, PESIT Bangalore South Campus, Bangalore, India
e-mail: anirvanmaiti@gmail.com

H. Veeradhi
e-mail: hema.veeradhi@gmail.com

S. Saha
e-mail: snehanshusaha@pes.edu

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_13

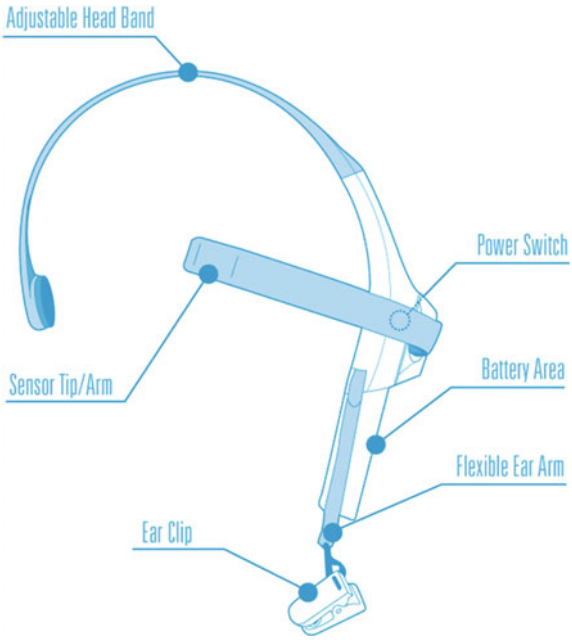


Fig. 1 NeuroSky device. Source https://www.support.neurosky.com/kb/mindwave/mindwave-diagram

online user experience while shopping must be recreated in such a way that it should match shopping in person. By analyzing the customers' reactions and responses, we can design the site accordingly and give a positive feedback to both the user and the e-commerce merchant.

EEG signal analysis has been gaining a lot of traction over the years with sophisticated devices such as Scalp-Cap with 64 electrodes capturing different bands of EEG signals [4–6]. Scalp-Cap, although very elaborate in terms of capturing signals, is an expensive device. We use the NeuroSky device which is a cheaper and viable option for the intended job. NeuroSky captures the user's attention, meditation, and eye blink strength levels for analyzing the reactions of the users while shopping. This device comprises two electrodes. One electrode is placed on the forehead and the other is connected to the earlobe, as shown in Fig. 1. The data recorded by the device is extracted in the form of a CSV file which is further processed and analyzed to extract relevant features, discussed in the later sections of the paper.

2 Literature Survey

EEG devices are used to measure a person's attention level in a number of different contexts [3], such as gaming and learning applications. Different methods are used to track a user's web browsing sessions. Some of the methods include eye gaze tracking, cursor movement, and scrolling [1]. Among these methods, eye gazing and scrolling

the web page in synchronization with the eye gaze caused discomfort to many users. Another method involves matching user's attention level to a particular section of the web page he/she was browsing, based on the scrolling position [2]. This helps in identifying the section of the web page a user is interested in. However, this may not be performed progressively, as the data depends on a specific time stamp on a particular day. As online shopping requires the need of analyzing customer behavior appropriately, our quantitative approach is more accurate and precise in predicting the user's actions. With the attention level captured from the NeuroSky Mindwave device, we monitor its fluctuations and observe the patterns as the user shops on the website. Website designers may implement the proposed system to enhance their website's performance-boosting business, therefore.

3 Proposed Approach

We first collect the data from the NeuroSky device through a third-party Android application called "EEG Analyzer". Data is gathered in the form of CSV files through this particular application. We store these files in our own DropBox folder. The data sets are then filtered and processed further in R.

The data sets include the time interval values (in milliseconds) and the corresponding attention, meditation, and eye blink strength values. We then try to identify the different user actions, i.e., going back to a previous page and buying and completing a transaction. Once feature identification is completed, a binary classifier model is then implemented to classify these features. Certain samples of the data sets are used for training the classifier, and the remaining is used for testing. The accuracy of the model is then calculated to verify its performance. The classifier model could

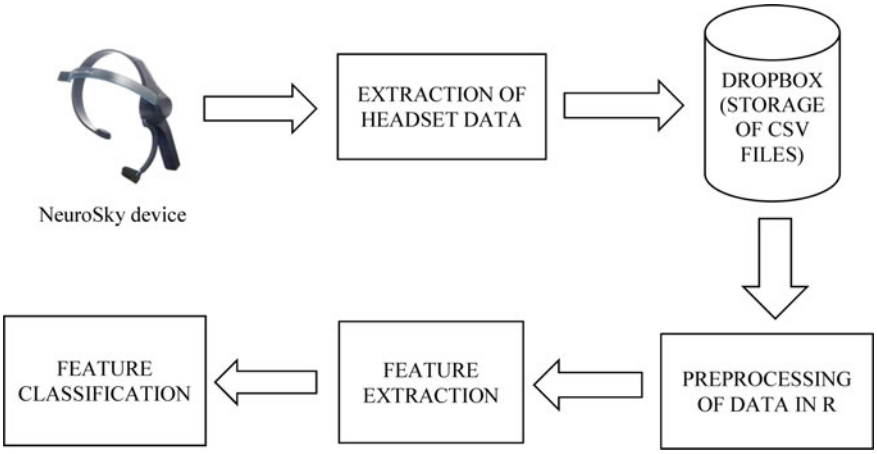


Fig. 2 System architecture

be handy to companies in order to analyze customer's reactions and enhance their website performance accordingly (Fig. 2).

4 Implementation

The experiment is conducted in two stages as follows:

4.1 Stage 1: Data Collection, Preprocessing, and Classification

4.1.1 Data Collection

In this module, we collect the data from the NeuroSky device of four subjects shopping on three different reputed e-commerce websites. The subject wears the device while he/she shops online and is disconnected after they complete shopping. A split timer is used to record the time during which a user performs the following actions:

- Moving to a new page.
- Scrolling up/down on a page.
- Applying product filters such as price, color, brand, etc.
- Clicks on Quickview—a feature to amplify the image of a product.

4.1.2 Preprocessing

After the sample data sets are collected, we then filtered all the unwanted information. Since the attention levels of the user are required, we eliminated the meditation and eye blink strength values (Type 5 and 22 values are removed from the CSV file). The sample data set contains the time in milliseconds denoted in scientific notation and is converted into shorter integer values for easier understanding. Null values from the data set are also eliminated.

4.1.3 Decision Tree Classification

After preprocessing the sample data sets, a graph of the time versus attention level is plotted as shown in Fig. 3. The time intervals during which the user performs various actions while shopping are marked (these time interval values were recorded by us during the data collection phase). The next phase involves identifying the time period a user would go back to a previous page based on the attention level values. During the data collection phase, we have also noted the time of user's return to a previous

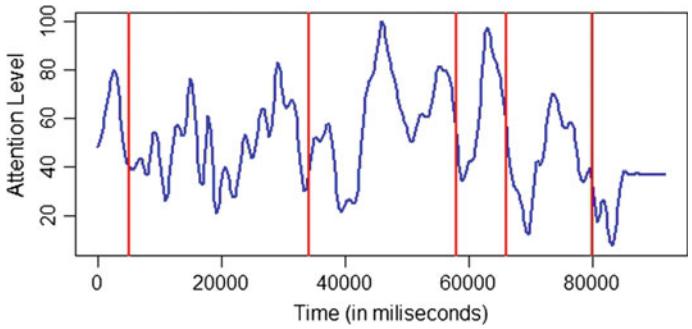


Fig. 3 Graph of time versus attention level (Website a data set)

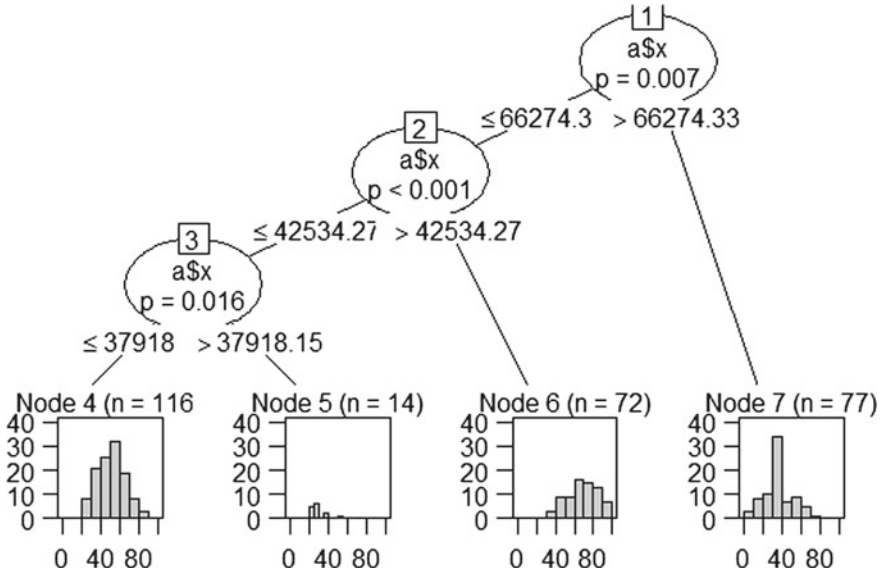


Fig. 4 Decision tree

page, transaction, and exit from the site. It was observed that the intervals during which such actions are performed occur near/at the local maxima and minima of the curve plotted.

Based on this empirical observation repeated many times, we then constructed a decision tree based on the *x*-axis, i.e., time, as shown in Fig. 4. The decision tree divides all the data points lying on the curve into different terminal nodes. The data points are distributed among the nodes depending on where they lie with respect to the maxima and minima of the curve. As we know, the time during which the user performs various actions, we may traverse the tree, following the branches based on the time values, to reach a terminal node. This terminal node represents the number

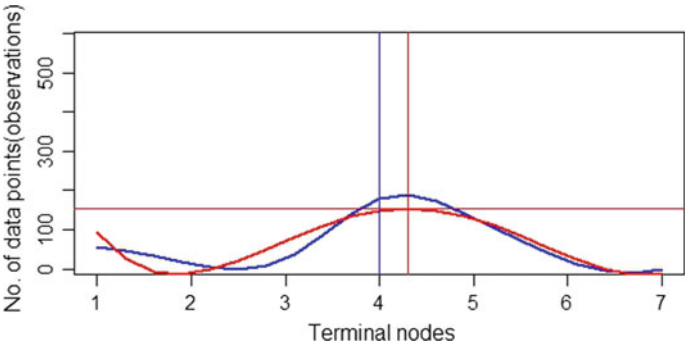


Fig. 5 Graph of terminal nodes versus number of data points present at each node for website's data set

of data points present at that node and their corresponding attention level values. After traversing the tree, it is easy to observe that the terminal node corresponding to the point in time where the user goes back to a previous page had the maximum number of data points (hence, it occurs at maxima).

A new graph depicting the number of terminal nodes present in the tree versus the number of data points present at each node is plotted, as shown in Fig. 5. As observed, a node corresponding to the user going back occurs at the global maxima of the curve (largest overall value of the curve) and the node corresponding to the user buying and exiting from the site occurs at the global minima (smallest overall value of the curve). Hence, by calculating the maxima and minima of this graph, we can approximate the time at which the user (i) goes back to a previous page and (ii) buys and exits the site.

4.2 Stage 2: Construction of Naïve Bayes Classifier

In this module, we implement a Naïve Bayes classifier based on the features identified from the previous module. The following features have been considered:

- User going back to a previous page;
- User buys, completes transaction, and exits the site;
- Other generic actions (scrolling, applying product filters, etc.).

The Naïve Bayes classifier checks to see whether the data points on the curve correspond to either the minima or the maxima. If it corresponds to the following:

1. Maxima—it labels it as belonging to the class "Going back to a previous page".
2. Minima and is present at the last terminal node—it labels it as belonging to the class "Buys and completes the transaction".
3. Else—it labels those points as belonging to the class "Generic actions".

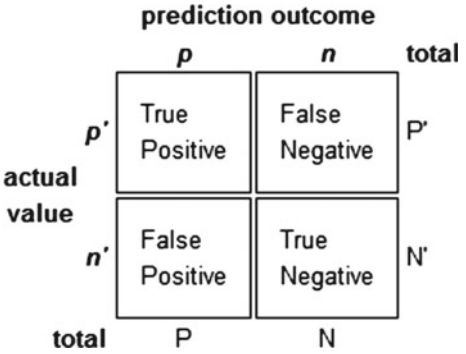


Fig. 6 Confusion matrix. Source https://www.tex.stackexchange.com/questions/20267/how-to-construct-a-confusion-matrix-in-latex

Naïve Bayes classifier is based on the Bayes theorem. The only difference is that Naïve Bayes classifier considers the predictors to be independent of each other. Given below is the equation of Bayes theorem:

$$p(C|x) = \frac{p(x|C)p(C)}{p(x)}, \tag{1}$$

where $p(C|x)$, called the posterior probability, is calculated from $p(C)$, $p(x|C)$ and $p(x)$.

Naïve Bayes classifier assumes that given a class c, the value of a predictor x is independent of other predictors. Thus, the posterior probability, $p(C|x)$, is computed as follows:

$$p(C|x) = p(x_1|C) \,.\, p(x_2|C).....p(x_n|C).p(C) \tag{2}$$

Next, in order to estimate the performance of the classifier model, a confusion matrix (also known as, Error matrix) has been used. A confusion matrix (shown in Fig. 6) demonstrates the number of accurate and inaccurate predictions made by the classifier as opposed to the real outcome from the data.

The Naïve Bayes classifier indicates the overall results in terms of sensitivity, specificity, accuracy, and precision which can be expressed as

$$sensitivity = \frac{TP}{TP + FN} \tag{3}$$

$$specificity = \frac{TN}{TN + FP} \tag{4}$$

$$accuracy = \frac{(TP + TN)}{TP + FP + TN + FN} \tag{5}$$

$$precision = \frac{TP}{TP + FP} \tag{6}$$

5 Results

The experiment was performed by using three different reputed e-commerce websites, with four different subjects. The experimental procedure was repeated for each of the website's data sets. The Naïve Bayes classifier implemented yielded an accuracy of 76, 88, and 93.75% for the three classes identified (i.e., Going back to a previous page, buying and exiting the site and other generic actions, respectively).

Every graph (time vs attention level) obtained is represented as a spline curve with an R-square value ranging between 85.14 and 87.80%. The snapshot of the superimposed graph of three different data sets of website A is given below in Fig. 8. The vertical lines indicate the node at which the user goes back to a previous page. As we can see, this node corresponds to/approximates to the maxima of the curve. The minima corresponding to the last terminal node represents the action of the user buying and exiting from the site. The curve colored in blue was used as the training data and the remaining data sets were used for testing purpose.

5.1 Splines

A spline function is generally represented as a piecewise function of the power of n. The points where the pieces connect are called knots, which provide the continuity and the smoothness to the function.

We used the spline functions, in order to achieve a higher degree of smoothness for the sudden fluctuations in the attention level of the user.

5.2 ROC (Receiver Operating Characteristic) Curve

ROC curve is a tool which is used to access the performance of a classifier model at all possible cut-off values. In this curve, the true positive rate (TPR) is plotted against the false positive rate (FPR) for various cut-off values. We have plotted the ROC curves for each of the three classes identified as follows (Figs. 7 and 8):

The spline equations for the curves are

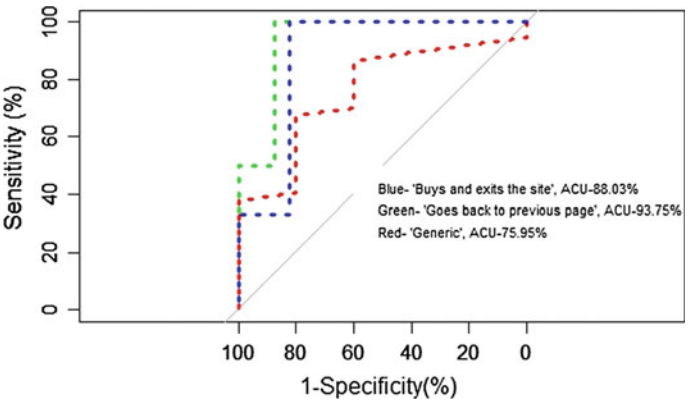


Fig. 7 ROC curve

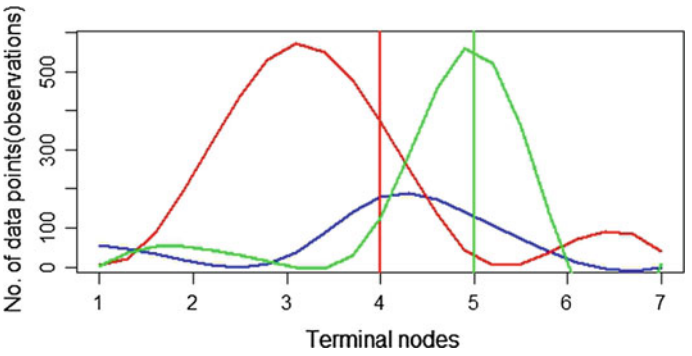


Fig. 8 Superimposed graph of terminal nodes versus number of data points present at each node

$$BlueCurve : -103.88x^4 - 185.30x^3 + 579.05x^2 - 441.81x + 94.08 = 0 \quad (7)$$

$$RedCurve : -120.89x^4 - 208.56x^3 + 446.63x^2 + 391.94x - 42.76 = 0 \quad (8)$$

$$GreenCurve : -1065.59x^4 + 3962.36x^3 - 3019.12x^2 + 1208.32x - 64.46 = 0 \quad (9)$$

6 Conclusion

We have evaluated the efficacy of the proposed system with a reasonable number of data sets obtained using four different subjects. Assuming that the user wears the NeuroSky throughout his/her online shopping duration and the EEG signals are correctly detected, the system is expected to work reasonably well with the chosen data sets.

The approach used to construct the system is considered to fit the current application. Many of the thresholds and constraints that are associated with the system were original and evolved through constant trial and error or plain intuitive thinking.

6.1 Future Scope

The features identified such as going back to a previous page, buying, and exiting from the site and other generic actions can be used by e-commerce website designers to make changes to their site in real time and dynamically.

Additional features such as scrolling, applying product filters, zooming in on a product, etc. can also be identified to make the online shopping experience more user-friendly.

Declaration Authors have obtained all ethical approvals from appropriate ethical committee and approval from the subjects involved in this study.

References

1. Beymer, D., Russell, D.M.: Webgazeanalyzer: a system for capturing and analyzing web reading behavior using eye gaze. CHI '05 Extended Abstracts on Human Factors in Computing Systems (2005). https://doi.org/10.1145/1056808.1057055
2. Bose, J., Singhai, A., Patankar, A., Kumar, A.: Attention Sensitive Web Browsing. Cornell University Library, Jan 2016. arXiv:1601.01092
3. Chen, C.-H., Tsai, T.-W.: Brain science approach to emotion analysis of web interface design. IASDR2013, Shibaura Institute of Technology, Tokyo (2013)
4. Kusuma, M., Saha, S., Srikantamurthy, K.: Brain-Computer Interfaces: Current Trends and Applications, volume 74 of Intelligent Systems Reference Library, chapter EEG Based Brain Computer Interface for Speech Communication: Principles and Applications, pp. 273–293 (2016). https://doi.org/10.1007/978-3-319-10978-7
5. Kusuma, M., Saha, S., Srikantamurthy, K.: Evidence of Chaos in EEG signals: An application to BCI, volume 337 of Studies in Fuzziness and Soft Computing, chapter Advances in Chaos Theory and Intelligent Control, pp. 609–625 (2016). https://doi.org/10.1007/978-3-319-30340-6_25
6. Kusuma, M., Saha, S., Srikantamurthy, K., Lingaraju, G.M.: Distinct Adoption of k-Nearest Neighbour and Support Vector Machine in Classifying EEG signals of Mental Tasks, volume 3 of Intelligent Engineering Informatics, pp. 313–329 (2015). https://doi.org/10.1504/IJIEI.2015.073064

Modified Rotation Forest Ensemble Classifier for Medical Diagnosis in Decision Support Systems

R. Ani, Jithu Jose, Manu Wilson and O. S. Deepa

Abstract Decision support system (DSS) in medical diagnosis helps medical practitioners in assessing disease risks. The machine learning algorithms prove a better accuracy in predicting and diagnosing diseases. In this study, rotation forest algorithm is being used to analyse the performance of the classifiers in medical diagnosis. The study shows that rotation forest ensemble algorithm with random forest as base classifier outperformed random forest algorithm. In this study, we use linear discriminant analysis (LDA) in place of PCA for feature projection in modified rotation forest ensemble method for classification. The experimental result also reveals that LDA can provide better performance with rotation forest while comparing with PCA. The accuracies given by random forest, rotation forest and proposed modified rotation forest classifiers are 89%, 93% and 95%, respectively.

Keywords Decision support system · Ensemble algorithm · Random forest Rotation forest · Principle component analysis · Linear discriminant analysis

R. Ani (✉) · J. Jose · M. Wilson
Department of Computer Science and Application, Amrita Vishwa Vidyapeetham, Amrita School of Engineering, Amrita University, Amritapuri, India
e-mail: anir@am.amrita.edu

J. Jose
e-mail: jithueruppakkattu@gmail.com

M. Wilson
e-mail: manuvilakunnathu@gmail.com

O. S. Deepa
Department of Computer Science and Application, Amrita Vishwa Vidyapeetham, Amrita School of Engineering, Amrita University, Coimbatore, India
e-mail: os_deepa@cb.amrita.edu

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_14

1 Introduction

Decision support system for medical diagnosis is a computer-based application which provides accurate and fast decisions based on the prediction algorithms applied on previous data from patients' records. The main idea behind the decision support system is to assist the medical practitioners at the point of care by analysing the medical data and recommend a method for treatment. In this study, the decision support system is trained using selected classification algorithm.

Coronary artery disease is one of the most common causes of death among the people. Heart disease occurs when a deposit of fat content accumulates at the coronary arteries. The deposit of fats inside the artery walls results in less flow of blood inside the arteries. The deposit hardens the walls of arteries and narrows down the path for the flow of blood. Risk assessment tools are widely used to assess the risk levels for a future heart disease. Recent research in the applications of data mining and machine learning in medical diagnosis helps to find out the common factors that are used to identify the cause of heart disease. The accuracies of the decision support systems for heart diseases play an important role in reducing the death rate due to heart failure. There are different symptoms and physiological changes related to cardiac diseases. The diabetes, hypertension, high blood pressure and high cholesterol are noticed as the main risk factors of heart diseases. There are other risk factors such as eating habits, obesity, physically inactivity, alcohol consumption and smoking associated with the heart disease.

Advancement in data mining applications in medical diagnosis is a boon to the modern society. The analyses and prediction in case of medical data can be done with the help of data mining technique such as complex algorithm. Accurate diagnosis at the early stage and subsequent treatment reduce the mortality rate due to diseases. Different data mining techniques are used for developing decision support system for early diagnosis of diseases such as heart attack, kidney failure, diabetes and stroke.

Classification algorithms are supervised learning technique in the field of data mining. In the classification of training data set, D consists of set of data records $D = \{X1, X_2, \ldots, X_N)$, and each data has a class label from the set of labels $Y = \{1, 2, \ldots, k)$. A classification model is being made from the training sample based on any of the classification algorithms (Fig. 1).

Ensemble classifiers mean a group of individual classifiers which are combined to form a new predictive model. The basic goal of ensemble classifier is to create more precise, accurate system. By using ensemble classifier, we can compensate the possibility for error that may occur by one base classifier. Different ensemble classifiers are created to achieve diversity among the base classifiers. Ensemble algorithm creates a set of classifiers, and then, prediction for new data points is done by the majority voting from all weak classifiers. Many of the ensemble algorithms use decision tree as the base learner.

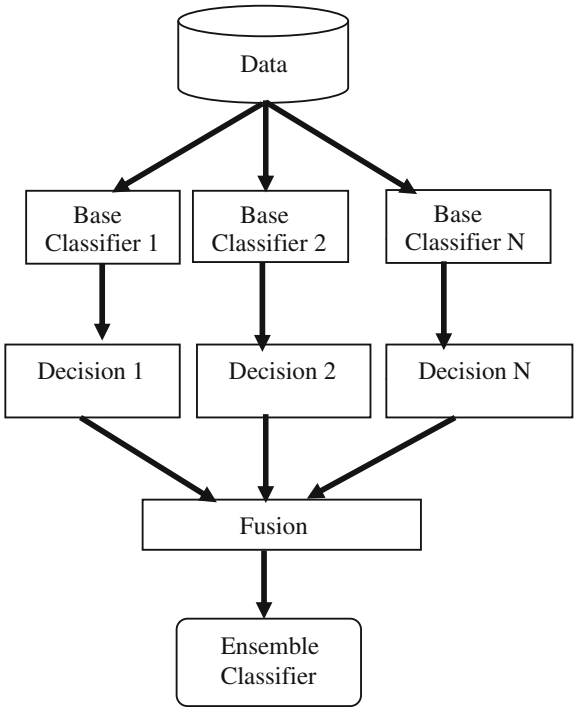


Fig. 1 Ensemble of classifiers

2 Related Works

This section of the paper explains the data mining techniques and researches that are already done for the prediction of heart disease.

2.1 In this paper [1], they use rotation forest (RF) with different base classifiers around 30 machine learning techniques to find the classification accuracy in Parkinson's disease, diabetes data set and heart diseases. Leave-one-out validation technique is used for all the experiments. The different classifiers show 77.42%, 72.25% and 84.53% accuracy for diabetes disease data set, heart disease data set and Parkinson's disease data sets, respectively. The rotation forest classifier ensembles have an average accuracy of 74.47%, 80.49% and 87.13% for diabetes disease data set, heart disease data set and Parkinson's disease data sets.

2.2 In this study [2], they use rotation forest algorithm with principal component analysis as an ensemble classifier model and also use artificial neural network in the role of base classifier. The newly proposed rotation forest with ANN model is tested with Wisconsin breast cancer data set taken from UCI repository. The final result shows that rotation forest with ANN structure was successful than particle swarm optimization with ANN structure. Rotation forest with ANN classifies breast cancer data set with an accuracy of 98.05%.

2.3 'Cancer classification using Rotation Forest' [3]. Here, the transformation method used is independent component analysis (ICA) instead of PCA. The efficiency of rotation forest classifier was checked by using breast cancer and prostate data set. The result shows that rotation forest gives better classification accuracy for microarray and also gives better accuracy, mainly in case of small ensemble size. Final result shows that rotation forest with ICA is much efficient than that of PCA.

3 Methodology

3.1 Random Forest

Random forest algorithm is an efficient prediction tool in machine learning. Using the bagging technique, a random subset of training data sets for every tree can be generated. For bagging technique, various extracted subsets are chosen. For every subset, a tree is created based on the splitting criteria. At last, it creates N number of trees based on different subsets. Random forest algorithm will predict an unknown class by considering the predicted values from all N trees and gives output based on majority voting technique.

3.2 Rotation Forest

The rotation forest tries to increase the diversity of individual decision tree by applying a rotation transformation matrix to the training data set before the training of each individual decision tree. The main analysis about the algorithm is to apply feature reduction technique and recreate a new feature set for every classifier in the ensemble. This can be done by splitting the feature set into k subset randomly and apply principal component analysis (PCA) to every created subset separately, and then by combining principle components of each subset, a new feature set is obtained. In this study, the base classifier for rotation forest is random forest. The rotation forest algorithm is explained by Rodriguez et al. [4] as follows.

Let X can be the sample of training data set. Let Y be the vector with class label $Y = [y_1 \ldots y_n]$. Let L be the number of base classifiers $C_1 \ldots C_L$ in the rotation forest. The following steps should be processed for every base classifier C_i:

Step 1: First, split the feature set F into K subsets. The subset can be disjoint or intersecting. In order to obtain high diversity, disjoint subset can be selected.

Step 2: For every subset, a non-empty subset of features is selected randomly, and then select 80% of data for training. Then, run PCA individually on each of the feature set $F_{(i,j)}$ that contains only the M features. Arrange the coefficients obtained from principal components analysis in the rotation matrix.

Step 3: Creating rotation matrix, R_i

$$R_i = \begin{bmatrix} a_{i1}^{(1)}, \ldots, a_{i1}^{(M_1)} & [0] & \ldots & [0] \\ [0] & a_{i2}^{(1)}, \ldots, a_{i2}^{(M_2)} & \ldots & [0] \\ \ldots & \ldots & \ldots & \ldots \\ [0] & [0] & \ldots & a_{ik}^{(1)}, \ldots, a_{ik}^{(M_k)} \end{bmatrix}$$

Step 4: Creating rearranged matrix.

The actual rotation matrix R_{ai} can be created by rearranging R_i to the feature set of data set X. After creating R_{ai}, X is multiplied with R_{ai} so that XR_{ai} is created. In order to classify a new input, x, the confidence of every class is predicted as

$$\mu_j(x) = \frac{1}{L} \sum_{i=1}^{L} dy\left(x^{'}\right), \quad j = 1, \ldots, c.$$

Assign x in the class having highest confidence.

Principle Component Analysis (*PCA*)

Principal component analysis (PCA) is used to bring out strong pattern and variations among features. It makes the data clear and easy to explore. Orthogonal projection method is used in PCA in order to convert the values in correlated variables into uncorrelated variables, which are known as principal components. The total count of principal components will be less than or same as that of original data set. PCA is a tool which is mainly used for data analysis and for creating predictive model. It also reduces the number of dimensions, without any loss in the information.

3.3 Modified Rotation Forest

In this study, modifications are made in the feature extraction part before finding out the rotation matrix, and an ensemble tree classifier is also chosen for the modified rotation forest. In modified rotation forest, linear discriminant analysis (LDA) is used instead of principal component analysis (PCA). The base classifier selected is random forest instead of simple decision tree.

Linear Discriminant Analysis (*LDA*)

Linear discriminant analysis (LDA) is a linear transformation method used to find out the linear combinations of features that provide best possible separation among the groups in original data set. LDA method is mainly used for pattern recognition, statistics and machine learning. While comparing LDA and PCA, LDA is supervised while PCA is unsupervised learning. PCA is called unsupervised because it avoids the class labels, and it targets only in the directions that maximise the

variance in the data. LDA computes the directions which represent axes that maximise the separation between multiple classes. LDA and PCA are used for analysing the data set in a different projection. The aim is to project data onto a lower dimensional space with good class-separability. Steps in linear discriminant analysis are as follows:

- Divide the feature set on the basis of number of classes.
- Calculate the mean of each attribute in different groups.
- Calculate the global mean of the feature set.
- The global mean is subtracted from each value in the data set.
- Covariance matrix is calculated.
- Pooled covariance matrix is calculated.
- Find the inverse of pooled covariance matrix.

A rotation matrix R_i is then created using the coefficients obtained from the calculation of inverse of pooled covariance matrix and then XR_{ai} is obtained.

4 Data Set Details

The data set used in the study is the publicly available UCI heart disease data set. This data set consists of 14 attributes and 303 instances. The attributes are as follows: age, sex, chest pain, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal and class.

5 Result and Discussion

The result obtained for data set is measured with the help of following measures:

- True Positive (%TP): Positive tuple which is labelled correctly with the help of classifier.
- True Negative (%TP): Negative tuple which is predicted correctly with the help of classifier.
- False Positive (%FP): Negative tuple which is incorrectly labelled as positive.
- False Negative (%FN): Positive tuple that is incorrectly labelled as negative.

Table 1 The confusion matrix that is obtained from random forest classifiers for each class. Here, Zero and One represent class labels. Zero means patient without heart disease. One means patient with heart disease

Confusion matrix		Prediction	
		Zero	One
Model	Zero	147	21
	One	24	111

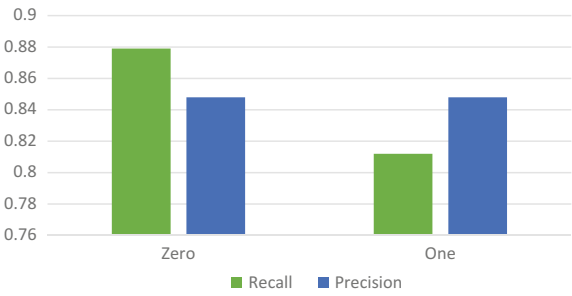


Fig. 2 The precision and recall obtained from random forest algorithm for each of the classes

5.1 Evaluation of Random Forest

See Table 1 and Fig. 2.

5.2 Evaluation of Rotation Forest

See Table 2 and Fig. 3.

5.3 Evaluation of Modified Rotation Forest

See Table 3 and Fig. 4.

Table 2 The confusion matrix that is obtained from rotation forest classifiers for each class. Here, Zero and One represent class labels. Zero means patient without heart disease. One means patient with heart disease

Confusion matrix		Prediction	
		Zero	One
Model	Zero	155	8
	One	12	128

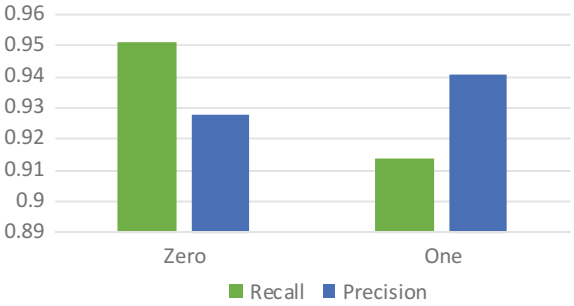
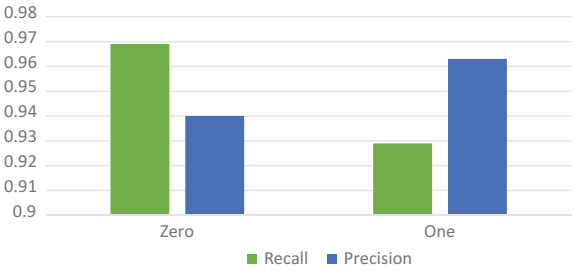


Fig. 3 The precision and recall obtained from rotation forest algorithm for each of the classes

Table 3 The confusion matrix that is obtained from modified rotation forest classifiers for each class. Here, Zero and One represent class labels. Zero means patient without heart disease. One means patient with heart disease

Confusion matrix		Prediction	
		Zero	One
Model	Zero	157	5
	One	10	131

Fig. 4 The precision and recall obtained from modified rotation forest algorithm for each of the classes



5.4 Accuracy Assessment

See Table 4 and Fig. 5.

Table 4 Comparison of base classifiers used in rotation forest and modified rotation forest

Base classifier	Rotation forest (%)	Modified rotation forest (%)
Random forest	93	95
Decision tree	82	83
LDA	75	75
Naïve Bayes	84	81

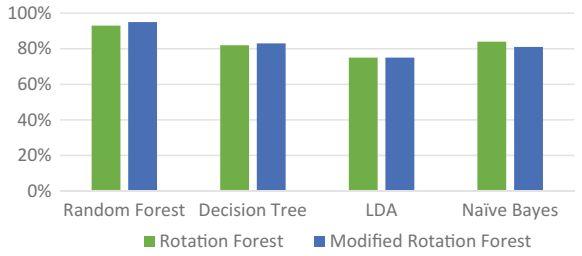
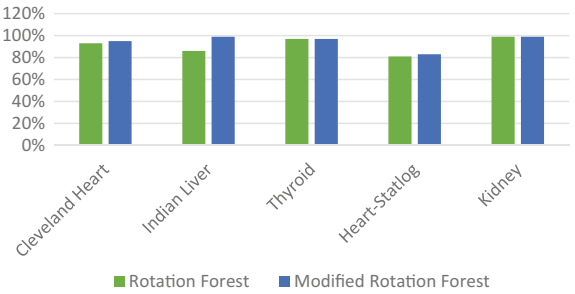


Fig. 5 The accuracy obtained from each base classifier

Table 5 Comparison of different data set with rotation forest and modified rotation forest

Data set	Rotation forest (%)	Modified rotation forest (%)
Cleveland heart disease	93	95
Indian liver patient	86	99
Thyroid	97	97
Heart-Statlog	81	83
Kidney	99	99

Fig. 6 The accuracy obtained for different data sets



5.5 Evaluation of Different Data Sets

See Table 5 and Fig. 6.

In this study, five different disease data sets are analysed using rotation forest and modified rotation forest. Almost all data sets provide better result in modified rotation forest while compared to rotation forest.

6 Conclusion

The proposed algorithm uses a better feature transformation technique on the data set, and the classification result shows an increase in accuracy when applied to different data sets. The proposed rotation forest algorithm uses random forest as the

base classifier. The analysis of performance of different base classifiers in rotation forest algorithms is done. The results proved better accuracy of an ensemble decision tree—random forest in proposed rotation forest. Future scope of modified rotation forest includes analysis of accuracy of algorithm for different numbers of feature subsets (k), and training of base classifier may be implemented in a parallel computing environment.

References

1. Ozcift, A., Gulten, A.: Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithm. Comput. Method Program Biomed. (2011)
2. Koyuncu, H., Ceylan, R.: Artificial neural network based on rotation forest for biomedical pattern classification. In: IEEE Conference 2013
3. Liu, K.-H., Huang, D.-S.: Cancer classification using rotation forest. Comput. Biol. Med. (2008)
4. Rodrıguez, J.J., Kuncheva, L.I., Alonso, C.J.: Rotation forest: a new ensemble classifier method. IEEE Trans. Pattern Anal. Mach. Intell. **28**(10) (2006)
5. Blaser, R., Fryzlewicz, P.: Random rotation ensembles. J. Mach. Learn. Res. **2** (2015)
6. Kuncheva, L.I., Rodríguez, J.J.: An Experimental Study on Rotation Forest Ensembles. Springer (2015)
7. Ozcift, A.: Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis. Comput. Biol. Med. **41** (2011)
8. Zhang, Z., Xie, X.: Research on AdaBoost.Ml with Random Forest", Conference on Computer Engineering and Information Technology, 2010
9. Krishnaiah, V., Srinivas, M., Narsimha, G., Subhash Chandra, N.: Diagnosis of Heart Disease Patients Using Fuzzy Classification Technique. IEEE (2012)
10. Karaolis, M., Moutiris, J.A., Pattichis, C.S.: Assessment of the risk of coronary heart event based on data mining. In: 8th IEEE International Conference on Bioinformatics 2008, pp. 1–5
11. Pavlopoulos, S.A., Stasis, A.Ch., Loukis, E.N.: A decision treebased method for the differential diagnosis of aortic stenosis from mitral regurgitation using heart sounds. Biomed. Eng. OnLine **3**, 21 (2004)
12. Rajeswari, K., Vaithiyanathan, V., Neelakantan, T.R.: Feature selection in ischemic heart disease identification using feed forward neural networks. In: International Symposium on Robotics and Intelligent Sensors 2012
13. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)

Social Data Analytics by Visualized Clustering Approach for Health Care

K. Rajendra Prasad, I. Surya Prabha, N. Rajasekhar and M. Rajasekhar Reddy

Abstract Social networks play a vital role in public healthcare systems. Twitter, Facebook, and blogs have millions of health-related content and it is required to filter the data for the reduction of processing cost. A semi-supervised health classifier model is proposed for health care which analyzes the patient condition by symptoms and recommends either suggestions or treatments for the relevant diseases such as influenza, flu, etc. In a proposed system, ailments clusters are defined based on the features of diseases using Visualized Clustering Approach (VCA). The proposed Twitter classifier model effectively works for high-rated risk diseases when compared to the traditional healthcare model. Results are discussed in the experimental study.

Keywords Social networks · Classifier · VCA · Ailments
Health care

1 Introduction

Social media is one of advances in Information Technology (IT) [1], which is used in many public sectors for different social applications. It is a prominent field in private and public healthcare sectors. In healthcare sector, it is required to properly mine and analyze a big volume of user-generated content, which improves the

K. Rajendra Prasad (✉) · I. Surya Prabha · N. Rajasekhar · M. Rajasekhar Reddy
Institute of Aeronautical Engineering, Hyderabad, India
e-mail: krprgm@gmail.com

I. Surya Prabha
e-mail: ipsurya17@gmail.com

N. Rajasekhar
e-mail: rajasekharnennuri@gmail.com

M. Rajasekhar Reddy
e-mail: mrajasekhar509@gmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_15

quality of services for public healthcare domain by optimizing the cost. The advantage of big social data analytics includes readily available bulk data, proper monitoring, and reduction in response time. In the social data analytics, novel visualized clustering approaches are proposed for the best services of health care in both private and public sectors.

Many methods [2] are used for analyzing the contents of social media networks. However, collecting the raw data, i.e., tag words regarding the content of health in a social network are the initial step in the research of social media based healthcare system. The next step is to perform the clustering for identifying specific health stream and its sub-domain using proposed clustering methodologies. After defining the health clusters (or streams), model is defined for each and every health clusters. Suggestions and solutions for curing a disease in any health stream are collected from social media analytics; these are incorporated using big social media techniques for the servicing of different health sectors.

The fundamental analysis of social media for healthcare system is divided into four key steps. In the first task, the data is collected from various social networks. The second task is breaking down the sentence into popular words and phrases on any required topic (such as insulin, strips, etc.). For this task, lexicon for natural language processing is required to develop an effective treatment. In this regard, the third task performs searching operation for identifying the keywords and links for annotating the products and services in healthcare system. In final task, effective visualized clustering approach (VCA) is developed for searching of patterns that are related to responses and product services of client's health care. For building social media based healthcare system, it is required to understand key consumer dependencies and maintain the information based on social network data. Healthcare system can also evaluate the level of patient's satisfaction through feedback of doctors from respective peers for the improvement of treatments.

Several studies have used social media for tracking health-related words, disease trends, monitoring of diseases, prevention of disease, and curing solutions of disease. Twitter is one such online data source, which consists of over a billion user tweet messages related to health topic.

Section 2 presents related work for healthcare system, Sect. 3 discusses two models of healthcare system, Sect. 4 presents experimental study of the healthcare system, and Sect. 5 presents conclusion and the future scope.

2 Related Work of Healthcare System

In the field of healthcare system, social media is the best resource for learning about people's health problems. Influenza is mostly commonly identified in social media. Linear regression [3–5], supervised classification [6, 7], and social media network analytics are some of the techniques used by researchers for detecting influenza on twitter. Social media is used for studying cholera [8], dental pain [9], cardiac arrest

[10], people behavior using physical activity [11], mood and mental health [12], and drug use [13]. In the existing work [14], performing discovery of ailments and health topics are described for the purpose of characterizing and identifying health problems through social media. However, it is necessary to build and strengthen the model for illness discovery and validate the illness using a prior knowledge of respective health clusters. Recent research of health care demonstrates that Twitter can provide more amount of information for various infection rates than that provided by the traditional techniques [15]. Twitter can be effectively used for disease surveillance. For these facts, Twitter model for various diseases is proposed in the following section.

3 Model for Social Media Healthcare System

By using people health information by social media, it is proposed to develop a prominent healthcare system. With this proposed system, we can analyze millions of tweets. In the proposed system, Ailment Topic Aspect Model (ATAM) [16] is used for analyzing millions of tweets. The following sequence steps are used for developing the social media healthcare system.

1. To describe the structure for data gathering and perusal of different data flow, filters, and supervised classifiers for recognizing the health records.
2. To examine millions of illness tweets using ATAM.
3. To separate and categorize non-ailment topics and ailment topics.
4. To classify the health problems using symptoms.
5. To detect health condition, propose solutions of preventions of illness.
6. To propose the treatment recommendation.

3.1 Latent Dirichlet Allocation (LDA) and Visualized Clustering Approach (VCA)

According to Latent Dirichlet Allocation (LDA) [17], words are distributed to form a topic and those topics are elaborated to form a text document. Some topics are not taken as input, even though the topic concludes. Respective topic models are unsupervised models, so these topic models will automatically cluster these words into topics and form respective documents. Preliminary LDA experiments discovered health-related topics around ailments.

Visualized clustering approach (VCA) [18] is used for assessment of health clusters through visual form, i.e., it shows the health clusters as square-shaped dark

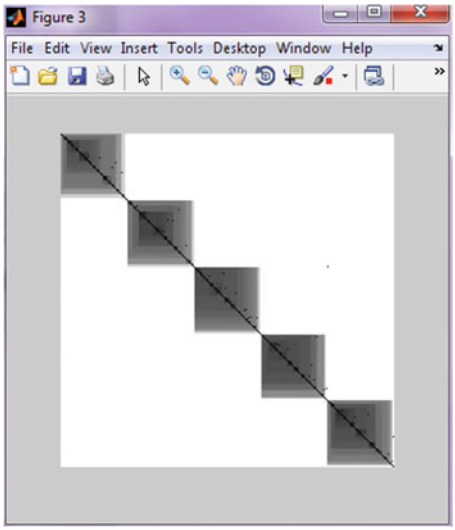


Fig. 1 Visual image of five health clusters

blocks along diagonal in a visual image, hence it is named as VCA. For example, Fig. 1 shows the image for understanding of square-shaped dark blocks as clusters.

The data related to five distinct health clusters is extracted from Twitter and is assessed after defining probabilistic models of health data, thus it is shown as five square-shaped dark blocks.

For example, some symptom terms belong to a topic cluster of different ailment. Consider the statement, "damn flu, home with a fever watching TV." In this, there are two words, pertinent to the ailment of flu ("flu" and "fever"), one of which is a symptom. Even though the statement did not belong to the illness.

A classier model is required to categorize illness from each tweet for distinguishing illness words from other topics.

3.2 Mining Trends

The proposed VCA is to discover groups of ailments from millions of tweets. In the proposed system, both VCA and classifier models are used for seeking extrinsic validation of groups for various health tasks. The treatment recommendations are received through classification model using prior knowledge. Extracting the tweet messages and mining the health trends, based on symptoms, is an ultimate aim of healthcare system for achieving public expected health-oriented outcomes.

4 Experimental Study of Healthcare System

For experimental study, millions of Twitter health-related data are extracted by data filtering. When two datasets from twitter of various time periods are taken, nearly 15,000 to 20,000 key phrases are collected that relate to illness, symptoms, and treatments through online websites. Some of the top words for ailments and health topics are shown in the following Table 1. The five categories of health data are summarized in Fig. 1, which shows five different square-shaped clusters.

The Pearson correlation is computed between the weekly influenza rate and monthly allergy rate in Twitter and same amount of influenza and allergy is measured from Centers of Disease Control and Prevention (CDC). These results are shown in Table 2.

It is noted that, twitter model is more correlated with CDC for flu specific health-related problems. Thus, social media provide prominent health treatment recommendations for influenza-like illness problems. In this experimental study, we have carried out five different issues related data and compared both Twitter model and CDC using Pearson correlation, and results are illustrated graphically through bar chart in Fig. 2. For flu and allergy cases, the correlation is high between Twitter and CDC. Hence, it is addressed that social media are prominent for flu and allergy cases in a public healthcare system. These social media models show the topics that can discover a significant ailment and confirm that Twitter model is strongly correlated with ground truth surveillance. Unsupervised VCA model automatically discovers the health-related word clusters and classifier models discover the ailments and treatments.

Table 1 Top words of Twitter health-related data

Ailments					
	Influenza-like Illness	Insomnia and Sleep Issued	Injuries and Pains	Dental Health	Diet and Exercise
General words	Ill	Asleep	Hurts, knee, leg	Dentist, tooth, teeth	Body, pounds
	Flu	Ill	Leg, arm	Apt, wisdom	Gym, weight
Symptoms	Sick, sore, throat, fever, cough	Sleep, headache, fall, insomnia,	Pain, sore	Infection, pain, mouth	Sore, aching
Treatments	Hospital, surgery, antibiotics, paracetamol, fluids	Sleeping, pills, caffeine, Tylenol	Massage, brace, therapy, crutches	Surgery, braces, antibiotics	Diet, exercise, protein

Table 2 Pearson correlation between Twitter model and CDC

	CDC (flu)	CDC (allergy)
Twitter model (flu)	0.652	Not applicable
Twitter model (allergy)	Not applicable	0.512

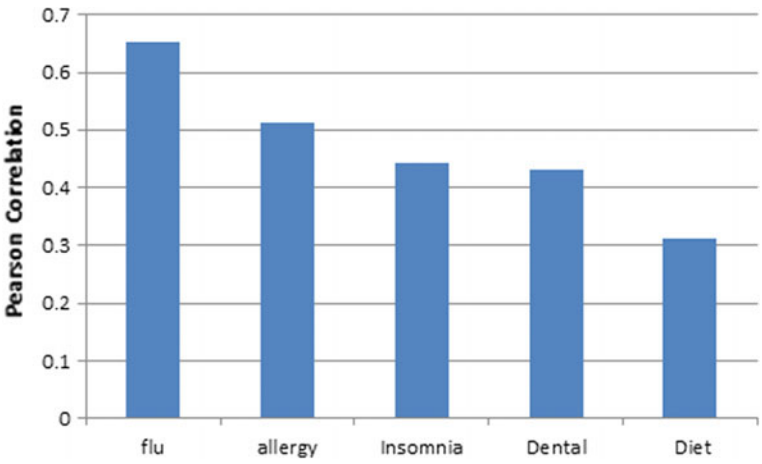


Fig. 2 Pearson correlation between Twitter model and CDC

5 Conclusion and Future Scope

This paper has addressed healthcare system using social media analytics. The proposed work is designed in a way of semi-supervised approach, i.e., it uses VCA for defining health clusters based on the symptoms of diseases and classier model is built for giving treatment labels or solutions for the diseases. Twitter model is evaluated for specific health cases and it is proved that social media networks are very useful for healthcare system. There is a large scope to extend healthcare system for addressing many challenging issues of social media networks.

References

1. Michael, J.P., Abeed, S., John, S.B., Azadeh, N., Matthew, S., Karen, L.S., Graciela, G.: Social media mining for public health monitoring and surveillance. In: Pacific Symposium on Biocomputing, pp. 468–479 (2016)
2. Batrinca, B., Philip, C.T.: Social media analytics: a survey of techniques, tools and platforms, AI & SOC **30**, 89–116 (2014)
3. Culotta, A.: Towards detecting influenza epidemics by analyzing Twitter messages. In: KDD Workshop on Social Media Analytics (2010)

4. Culotta, A.: Lightweight methods to estimate influenza rates and alcohol sales volume from Twitter messages. In: Language Resources and Evaluation. Special Issue on Analysis of Short Texts on the Web (2012)
5. Lampos, V., Cristianini, N.: Tracking the flu pandemic by monitoring the social web. In: IAPR 2nd Workshop on Cognitive Information Processing (2012)
6. Maskawa, S., Aramaki, E., Morita, M.: Twitter catches the flu: detecting influenza epidemics using Twitter. In: Conference on Empirical Methods in Natural Language Processing (2010)
7. Lamb, A., Paul, M.J., Dredze, M.: Separating fact from fear: tracking flu infections on Twitter. In: Conference of the North American Chapter of the Association for Computational Linguistics (2013)
8. Chunara, R., Andrews, J., Brownstein, J.: Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. Am. J. Trop. Med. Hyg. **86**, 1 (2012)
9. Heaivilin, N., Gerbert, B., Page, J., Gibbs, J.: Public health surveillance of dental pain via Twitter. J. Dent. Res. **90**, 9 (2011)
10. Bosley, J.C., Zhao, N.W., Hill, S., Shofer, F.S., Asch, D.A.: Decoding twitter: surveillance and trends for cardiac arrest and resuscitation communication. Resuscitation **84**, 2 (2013)
11. Yoon, S., Elhadad, N., Bakken, S.: A practical approach for content mining of tweets. Am. J. Prev. Med. **45**(1) (2013)
12. Golder, S., Macy, M.W.: Diurnal and seasonal mood varies with work, sleep and day length across diverse cultures. Science **333**(6051), 1878–1881 (2011)
13. Moreno, M., Christakis, D.A., Egan, K.G., Brockman, L.N., Becker, T.: Associations between displayed alcohol references on Facebook and problem drinking among college students. Arch. Pediatr. Adolesc. Med. (2011)
14. Michael, J.P., Mark, D.: Discovering health topics in social media using topic models. Plos one **9**(8), 1–14 (2014)
15. Blei, D.: Probabilistic topic models. Commun. ACM **55**(4), 77–84 (2012)
16. Michael, J.P., Mark, D.: A model for mining public health topics from twitter, Johns Hopkins University (2011)
17. Blei, D., Ng, A., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. (2003)
18. Rajendra Prasad, K., Eswara Reddy, B.: An efficient visualized clustering approach for various datasets. In: IEEE SPICES, NIT Calicut (2015)

Mining Efficient Rules for Scene Classification Using Human-Inspired Features

Padmavati Shrivastava, K. K. Bhoyar and A. S. Zadgaonkar

Abstract Researchers in the field of image understanding face the challenging issue of conceptual description at semantic level based on primitive visual features. The work presented in this paper attempts to provide semantic description of an image. The first objective of the proposed research is to extract features based on human perception from natural scene images. A mapping process is used to obtain high-level semantic features based on color, texture and structural features to tackle the problem of semantic gap. The second focus is to construct a classification model using classification rules mined from training images. The mined rules are easily understandable and have the advantage of supporting user's view of categorizing an image in a particular class. Two state of the art data mining algorithms: Classification based association (CBA) and decision tree induction are empirically evaluated for the purpose of classification of natural scenes. The results obtained by 10-fold cross-validation approach show that classification rules extracted using ID3 algorithm yield good performance with an average accuracy rate of 83.8%.

Keywords Image mining · Classification rules · Scene classification
Human-Inspired features · Decision tree · Semantic gap

P. Shrivastava (✉) · A. S. Zadgaonkar
Dr. C. V. Raman University, Bilaspur 495113, Chhattisgarh, India
e-mail: padmavati.shrivastava@yahoo.co.in

A. S. Zadgaonkar
e-mail: arunzad@gmail.com

K. K. Bhoyar
Yeshwantrao Chavan College of Engineering, Nagpur 441110, Maharashtra, India
e-mail: kkbhoyar@yahoo.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_16

1 Introduction

Due to technological advancements, the volume and type of digital images being generated have seen an enormous growth. From medical image collection to remote sensing images or a collection of vacation photographs, the heterogeneity and complexity of images makes image understanding by machines a challenging task. To overcome the time-consuming and cumbersome task of image analysis, it is essential to develop image interpretation systems based on machine learning. The dynamic characteristics of images require innovative approaches capable of visualizing data closely related to human perception. Image classification is an important application in the area of computer vision. It automatically assigns class label to a given image. Although image classification is a fundamental problem in multimedia research, it is difficult to describe the semantic meaning of an image based on its content or contextual information. A major research challenge is to assign a label to an image which is as close to human's perception as possible. Majority of the image classifiers are developed using low-level feature descriptions. The demand is that the classification decision should be defined in terms of semantic concepts rather than low-level visual cues. The research presented in this paper focuses on major issues which include transformation of low-level signatures to high-level semantics and extracting relevant knowledge in the form of rules useful in describing images.

2 Pattern Mining

Image mining is a research field which aims at discovering implicit relationships among patterns which may be hidden or not observable directly [1]. The knowledge discovered indicates associations and causalities which are helpful to generate high-level descriptions of image categories at different levels of user understanding. In large databases association, rule mining is used to determine relations between variables [2]. A popular data mining technique used to discover association rules and utilize them in building classification systems is Associative Classification. A pattern which establishes the relationship between occurrences of features and class labels is expressed in the form of an association rule. It is essential to integrate the knowledge of the application domain into the methodology used to make the rules mined from databases more useful. Several authors have applied the use of association rules to discover relationships based on various interestingness measures. In [3] authors propose two schemes for automatic image annotation using association rules. In the first scheme, color features are extracted for the entire image along with some textual descriptors and association rules are used to identify relationships between the two. The second approach uses local color information to derive clusters using k-means based clustering algorithm. To discover associations between features of a cluster, localized color, and textual descriptors are extracted

and a set of rules are derived for each cluster. In [4] Yin and Li propose a content-based image retrieval system based on association rules. Relevance feedback on the retrieved images is collected based on soft parameters. The number of association rules is decided based on confidence value along with measures to detect redundancy. Wang et al. [5] developed an image retrieval system based on the fusion of multiple modalities. Association rules have been used to establish correlations between visual features and image semantics. In [6] a semantic image retrieval system based on textual and visual features is suggested. Clustering and Association Rule algorithms which are two important data mining techniques are used to retrieve semantically similar images. Association rules are discovered between textual semantic clusters and visual clusters which are used later for the purpose of retrieval. Naik and Patel [7] have used the FP tree algorithm to discover association rules among the features extracted from a database of MRI images and the normal/abnormal category to which each image belongs. The discovered association rules are used to construct a classification system based on decision tree algorithm to identify and classify brain tumor as normal or one which belongs to malignant or benign class. Phadikar et al. [8] proposed an automation system in which different types of rice diseases are identified and classified. Firstly images of rice plant are used to identify the infected regions. Low-level features like color, shape and spatial location of the infected regions are then extracted. For dimensionality reduction, rough set theory is used to select discriminative features such that loss of information is minimized. The selected features are used to build rules such that all diseases pertaining to rice plant images are covered. Dash and Kolippakkam [9] have applied image mining for detection of Egeria which is a type of weed. The major challenge in detecting weeds in test images is to identify a set called view defined by appropriate features and their values. A view is identified using association rule mining technique. The best view is the one with highest support and confidence. Lucas et al. [10] have used temporal images which are captured at specific time intervals and extracted rules for the purpose of classification. The approach has been used to study the distribution of agricultural land and habitats. Ordonez and Omiecinski [11] addressed the problem of knowledge discovery and finding associations in image databases. The authors proposed a domain-independent system tested on synthetic image database of geometric shapes. In [12] Stanchev and Flint proposed an image retrieval system based on semantic features at a high level of description. Image mining technique has been utilized to derive fuzzy production rules for the conversion of low-level features into high-level semantic descriptions.

3 Association Rules

An association rule is used to discover patterns and links between data items. The two parts of an association rule are an antecedent (if) and a consequent (then) thus making it a conditional statement. To find rules in images, feature elements (either

numeric or categorical) are treated as data items. A number of methods exist to mine association rules from a database. The generated rules should uncover interesting patterns and must be simple, easily understandable, and strong. Support and confidence are two important measures used to determine the strength and interestingness of the discovered rules.

In a database containing a set of transactions, the fraction of transactions containing both the antecedent and the consequent in a rule is used to determine the *Support* value of the rule.

Similarly given a set of transactions, the *Confidence* value of a rule can be calculated by counting the number of transactions in which consequent occurs provided that the antecedent is true.

For a rule R: Antecedent $\rightarrow$ Consequent

Support (R) = Support (Antecedent $\cup$ Consequent)

Confidence (R) = Support (Antecedent $\cup$ Consequent)/Support (Antecedent)

Any association rule mining algorithm for a given set of transactions T aims to find all rules whose support value is at least equal to a user-defined minimum support threshold and confidence value is at least equal to a user-defined minimum confidence threshold. Careful selection of these two parameters is essential since support represents the statistical significance or prevalence of rules and confidence represents the goodness or predictability of rules. A high value of *support* may generate few rules whereas a low *support* value may discover a large number of rules many of which may be uninteresting. We now discuss two different algorithms for rule discovery:

CBA (Classification based Association) Algorithm:

Association rules were originally designed for finding correlations between items in transactions. The CBA algorithm [13] is used for class prediction and is based on mining frequent item sets. The purpose of CBA is to find Class Association Rules (CARs) of the form X $\rightarrow$ Y [If Body then Head] where X represents item sets and Y represents a class label. It implements the Apriori algorithm [14] where the threshold values for support and confidence measures are specified by the user and all association rules with their support and confidence measures at least equal to the threshold values are generated. When a test example is to be classified, the rule with highest confidence measure with matching body is chosen and the corresponding class label is assigned to the test example.

ID3 Decision Tree Algorithm

ID3 algorithm is used for building a *Classification Model* from data in the form of a *Decision Tree*. The decision tree comprises of different nodes and arcs connecting these nodes where each node corresponds to an attribute or a feature and each arc represents the value of that attribute. Every leaf represents a path starting from the root and specifies the class or category to which the data described by the path belongs. The decision of which attribute to be associated at a node depends on the information gain measured as entropy of the node. If we have c classes then we define entropy of a system S as

$$\text{Entropy(S)} = \sum_{i=1}^{c} p_i log_2 p_i,$$

where p_i is the probability of occurrence of class i.

The decision tree algorithm has been chosen because of the simplicity in deriving a set of rules.

4 Proposed System

The details of the approaches used to mine association rules based on human-inspired features which are closer to human perception from a collection of natural scenes are presented in this section. The system comprises of the following steps: Low-level Feature Extraction, Association Rule Mining, and Classification. The application domain chosen for the implementation of our approach is categorization of natural scenes. Natural scenes are complex since they are composed of few basic components which are arranged in an unpredictable manner. This makes their categorization a challenging task. Humans perceive scenes at a high level of concepts. For example, blue color and absence of objects in top part of the image with green textured region at the bottom indicates an open country image. We have used the Oliva-Torralba dataset (OT) [15] which is a subset of the Corel database in our experiments. For the purpose of scene classification, we have used only the four natural classes (coastal areas, forests, mountains and open-country-side) from the dataset. A total of 1472 images have been used for experimental evaluation. The proposed model organized at different levels is shown in Fig. 1.

4.1 Different Levels of Association Rules Based Classification Model

The proposed rule-based scene classification system using human-inspired features is organized into different levels as follows:

Level 0: Raw Pixels

This level comprises of image database consisting of natural scene images organized into four semantic categories. Each image is stored in jpeg format of the size 256 × 256 pixels. Every image pixel has the red, green, and blue components and is represented by a triplet.

Level 1: Primitive Features

At this level, primitive image features are extracted through different feature extraction methods and stored for further analysis. Various color, texture and edge

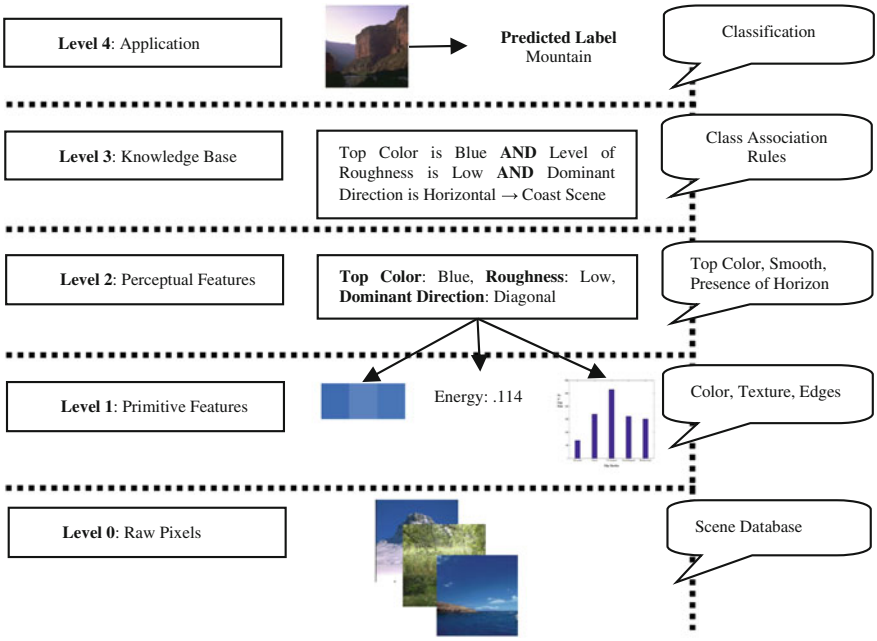


Fig. 1 Classification model based on association rules using human-inspired features

feature descriptors are utilized to represent image characteristics. Section 4.2 presents the detailed description of the features and their extraction mechanisms.

Level 2: Perceptual Features

The purpose of introducing this level between Level 1 and Level 3 is that human perception of scene-gist is far different from the image characteristics represented at low-level. For example at Level 1, foliage is recognized to have more entropy than water whereas human perception is that the two concepts differ in degree of roughness. The focus of Level 2 is to introduce human-inspired features which prove to be useful in reducing the semantic gap. A mapping process is applied to convert visual features at the primitive (lowest) level into high-level human-inspired features to generate semantically significant patterns. In Sect. 4.3, the details of the methods used are presented.

Level 3: Knowledge Base

The high-level semantic features obtained at Level 2 are subject to mining process to be able to uncover hidden correlations. The knowledge base is created in the form of Class Association Rules (CARs) to establish associations between semantic concepts and class label.

Level 4: Application Level

This is the highest level at which the rules generated using the entire training data are stored in the knowledge base and used for classification. The mined rules are then considered to categorize an unknown test image into one of the predefined classes.

4.2 Feature Extraction

A careful selection of domain-specific features is essential for any scene categorization system. Color, texture, and edges are intrinsic visual features which best contribute to differentiate among images from different scene categories. The features used in the proposed system are:

Naming Colors

Color is an important visual cue in scene perception. Each image is divided into three horizontal parts: *Top, Middle, and Bottom*. In order to obtain the dominant color of a region, the degree of membership of each image pixel in eleven basic categories of colors: Red, Green, Blue, White, Black, Gray, Pink, Orange, Yellow, Brown, and Purple is obtained using Fuzzy Color Naming Model presented in [16]. After mapping each image pixel to one of these color categories, the color which occurs most is considered as the color name of that region. This helps to obtain high-level color description of the three horizontal slices of the image.

Edge Direction

The Edge Direction Histogram keeps a count of the frequency of gradients of edges in user-specified directions. Edges are first detected using canny edge detector. The edge pixels in vertical, horizontal and two diagonal directions are counted. Additionally, nondirectional edges which do not lie along any of the above directions are also counted.

Horizon Line Detection

The presence of horizon line in an outdoor scene can be a discriminating feature which is useful in separating classes. A clear horizon line separating the sky is visible in most of the beach and countryside scenes. The image is decomposed into red, green and blue channels and its gradient is obtained. By applying Hough Transform, the longest line segment is detected. The slope of the longest line segment helps to ascertain if it corresponds to horizon line or not.

Texture description of top part of the image

Natural scenes are also categorized as open, semi-closed or closed. An open image has a vast sky in the top part of the image which is characterized by minimum or few objects and absence of texture. Hence texture descriptor of the top part of the

image serves as an important cue. In order to calculate texture characteristics, we decompose the top part of the image into ten equal-sized sub-blocks. The method outlined in [17] is then applied to each sub-block. The energy of each of the ten channels obtained after three-level wavelet decomposition [17] is calculated using the formula

$$C_n = \frac{1}{MN} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} |w(x, y)|,$$

where each channel is of dimension M by N and w(x, y) is a wavelet coefficient within the channel.

Statistical Texture Features

The amount of texture contained in an image is statistically determined by computing its Gray Level Co-occurrence Matrices (GLCM). Energy and Entropy are computed from GLCM for four different angles [0°, 45°, 90°, 135°] and displacement offset 2 to obtain texture features [18]. The average of GLCM energy obtained for each of the offset angle pair is used as a feature. Similarly, the mean of the GLCM entropy for each of the offset angle pair is considered as a texture feature.

4.3 Mapping of Primitive Features to Semantic Features

The mapping process is presented below and summarized in Table 1:

Naming Colors: Using the method already presented in Sect. 4.2, the top and bottom horizontal slice of the image is obtained and their dominant colors are assigned a name.

Dominant Direction: The dominant direction is one along which the number of edge pixels is maximum. Base on the dominant direction each region is assigned

Table 1 Mapping process to derive semantic features from low-level descriptors

Low-Level feature	Semantic feature	Mapping process
RGB values of each pixel	Color name	TSE color naming [16]
Edge count in five directions	Dominant direction	Horizontal, vertical, forward/backward diagonal, nondirectional
Horizon line detection	Presence/absence of horizon	Presence(Yes) absence(No)
Texture of top part	Textured/non-textured region	Low, medium, high [17]
GLCM entropy	Roughness	Smooth, medium, coarse
GLCM energy	Amount of texture	Low, high

one of the five labels—Horizontal, Vertical, Forward Diagonal, Backward Diagonal and Nondirectional.

Texture description of top part of the image: After computing the energy of each channel of the 10 sub-blocks of the top part of the image, the following factor as outlined in [17] is computed for each sub-block

$$R = \frac{C_1 + C_2 + C_3 + C_4}{C_5 + C_6 + C_7},$$

where Ci refers to the energy in the ith channel of an image sub-block.

A sub-block can be labeled as smooth if the factor R is greater than or equal to a T and it is textured if it is less than the T, where T is a threshold value to be determined experimentally. The top part of the image is labeled as follows:

Number of smooth regions is greater than 7: Amount of texture is 'Low'
Number of smooth regions is between 3 and 6: Amount of texture is 'Medium'
Number of smooth regions is less than 3: Amount of texture is 'High'

Horizon Line Detection: The presence of horizon in an image is labeled as 'Yes'and its absence is indicated by 'No'

Average GLCM Energy: Discretization process is followed to obtain a semantic label corresponding to GLCM Energy as follows:

Less than 0.2: 'Low'
Greater than 0.2: 'High'

Average GLCM Entropy: It is used to estimate the Level of Roughness in an image. To obtain a semantic label the following discretization process is used:

Less than 2.4: 'Low'
Between 2.4 and 3.2: 'Medium'
Greater than 3.2: 'High'

5 Experimental Evaluation

The mapping process results in a set of semantic features which are submitted to the mining process. Class Association rules are obtained by CBA and ID3 algorithms using Keel Suite which is a Java-based software tool for knowledge extraction. Sample association rules for few categories obtained by ID3 algorithm are given below:

If Dominant Direction = 'Horizontal' AND Top Color = 'Gray' AND Bottom Color = 'Green' THEN Class = 'Open Country'

Table 2 Results of classification using cross-validation approach (10-fold)

Fold number	ID3 algorithm	CBA algorithm
Fold 1	84.35	71.08
Fold 2	82.99	69.7
Fold 3	85.03	71.8
Fold 4	82.99	71.08
Fold 5	84.3	72.0
Fold 6	82.3	67.34
Fold 7	87.4	73.08
Fold 8	84.69	70.06
Fold 9	81.9	66.89
Fold 10	82.3	68.4
Average	83.825	70.01

If Top Color = 'Blue' AND Dominant Direction = 'Forward Diagonal' AND Bottom Color = 'Green' AND Level of Roughness = 'Smooth' AND Horizon Line = 'Yes' THEN Class = 'Mountain'

If Presence of Texture = 'High' AND Top Color = 'Green' AND Horizon Line = 'No' AND GLCM Energy = 'High' AND Bottom Color = 'Green' THEN Class = 'Forest'

The association rules obtained either by CBA algorithm or ID3 algorithm are used for prediction. To evaluate the system, a 10-fold cross-validation approach is used with 80% of the data used for training and remaining 20% for the purpose of testing. The classification results are given below (Table 2):

6 Conclusion and Future Directions

In this paper, a novel approach for natural scene classification based on semantic features is presented. The semantic gap problem is addressed with the use of features close to human perception. The other focus of this work is the development of a knowledge base by mining association rules. The proposed system has been applied to classify four categories of natural landscape scenes and good accuracy has been obtained. The results show the efficiency of the system. The advantage of the system is that the algorithms and features extracted from images at the lowest level are general. This allows the system to be suitable for other natural scene categories since it can be easily extended by generating association rules for new classes. In future work, other discretization methods may be implemented to overcome the loss of information due to mapping process. The present work can be extended to include semantic concepts such as sky and presence of certain objects to obtain rules.

References

1. Zhang, J., Hsu, W., Lee, M.L.: Image mining: issues, frameworks and techniques. In: Proceedings of the 2nd ACM SIGKDD International Workshop on Multimedia Data Mining (MDM/KDD'01). University of Alberta (2001)
2. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, DC, pp. 207–216 (1993)
3. Sethi, I.K., Coman, I.L., Stan, D.: Mining association rules between low-level image features and high-level concepts. In: Aerospace/Defense Sensing, Simulation, and Controls, International Society for Optics and Photonics, pp. 279–290 (2001)
4. Yin, P.Y., Li, S.H.: Content-based image retrieval using association rule mining with soft relevance feedback. J. Vis. Commun. Image Represent. **17**(5), 1108–1125 (2006)
5. Wang, Q., Lv, Y., Song, L.: Multimode Retrieval of Breast Masses Based on Association Rules. Int. J. Hybrid Inf. Technol. **8**(2), 389–396 (2015)
6. Alghamdi, R.A., Taileb, M.: Towards semantic image retrieval using multimodal fusion with association rules mining. In: Human Interface and the Management of Information. Information and Knowledge Design and Evaluation, pp. 407–418. Springer International Publishing (2014)
7. Naik, J., Patel, S.: Tumor detection and classification using decision tree in brain MRI. Int. J. Comput. Sci. Netw. Secur. (IJCSNS) **14**(6), 87 (2014)
8. Phadikar, S., Sil, J., Das, A.K.: Rice diseases classification using feature selection and rule generation techniques. Comput. Electron. Agr. Elsevier **90**, 76–85 (2013)
9. Dash, M., Kolippakkam, D.: Automatic view selection: an application to image mining. In: Advances in Knowledge Discovery and Data Mining, pp. 107–113. Springer, Berlin (2005)
10. Lucas, R., Rowlands, A., Brown, A., Keyworth, S., Bunting, P.: Rule-based classification of multi-temporal satellite imagery for habitat and agricultural land cover mapping. ISPRS J. Photogramm. Remote Sens. Elsevier **62**(3), 165–185 (2007)
11. Ordonez, C., Omiecinski, E.: Discovering association rules based on image content. In: Proceedings of th IEEE Forum on IEEE Research and Technology Advances in Digital Libraries, pp. 38–49 (1999)
12. Stanchev, P., Flint, M.: Using image mining for image retrieval. In: Proceedings of the IASTED Conference Computer Science and Technology, pp. 214–218 (2003)
13. Liu, B., Hsu, W., Ma, Y.: Integrating classification and association rule mining. In: Knowledge Discovery and Data Mining, pp. 80–86 (1998)
14. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In Bocca, J.B., Jarke, M., Zaniolo, C. (eds.) Proceedings of the 20th International Conference on Very Large Databases, Santiago, Chile, pp. 478–499 (1994)
15. Oliva, A., Torralba, A.B., Guerin-Dugue, A., Herault, J.: Global semantic classification of scenes using power spectrum templates. Challenge of Image Retrieval (CIR99), Electrical Work. In: Computing Series, Springer, Newcastle (1999)
16. Benavente, R., Vanrell, M., Baldrich, R.: Parametric fuzzy sets for automatic color naming. J. Opt. Soc. Am. A **31**(1), 48–56 (2008)
17. Porter, R.M.S., Canagarajah, C.N.: A robust automatic clustering scheme for image segmentation using wavelets. IEEE Trans. Image Process. **5**(4), 662–665 (1996)
18. Haralick, R.M., Shanmugam, K., Dinstein, I.H.: Textural features for image classification. IEEE Trans. Syst. Man Cybern. **3**(6), 610–621 (1973)

Patent Document Clustering Using Dimensionality Reduction

K. Girthana and S. Swamynathan

Abstract Patents are a type of intellectual property rights that provide exclusive rights to the invention. Whenever there is a novelty or an invention, prior art search on patents is carried out to check the degree of innovation. Clustering is used to group the relevant documents of prior art search to gain insights about the patent document. The patent documents represent hundreds of features (words extracted from the title and abstract fields). The common sets of features between the documents are subtle. Therefore, the number of features for clustering increases drastically. This leads to the curse of dimensionality. Hence, in this work, dimensionality reduction techniques such as PCA and SVD are employed to compare and analyze the quality of clusters formed from the Google patent documents. This comparative analysis was performed by considering title, abstract, and classification code fields of the patent document. Classification code information was used to decide the number of clusters.

Keywords Prior art search · Dimensionality reduction · Clustering

1 Introduction

Patents provide an excellent source of information on inventions and display the recent trends in technology. Prior art search is undertaken to ascertain whether the invention is novel, nonobvious and useful or not. It uncovers any knowledge that exists before the filing of the current invention. This knowledge includes patent

K. Girthana (✉) · S. Swamynathan
Department of Information Science and Technology, Anna University,
Chennai 600025, India
e-mail: k.girthana@auist.net

S. Swamynathan
e-mail: swamyns@annauniv.edu

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_17

documents, articles, books, conference proceedings, press releases, industry standards, and other related documents.

Over the recent years, the volume of filings and the capacity of IT systems to store and process had grown proportionately. At the same time, the growth of Internet has changed the public expectations and indeed the needs of the patent examiners particularly on prior art searching. Clustering is employed on this prior art search to retrieve relevant documents because the patent document may belong to multiple domains. Grouping the documents into clusters helps to invalidate the given patent application.

Though the structure of the patent documents is same, these documents are very verbose, and each document follows its lexicon as suggested by the patent document writer or the inventor. So, the number of common terminologies between the documents becomes very less. This causes sparseness in the document-term matrix (DTM) and clustering these data reduces the performance. Dimensionality reduction becomes a solution to the problem. The resultant reduced set of features is further processed. Many dimensionality reduction techniques are available in the field of statistics and machine learning [1, 2]. These techniques are performed through either Feature Selection or Feature Extraction.

Feature selection selects only the relevant and useful set of features from the original set and creates a new feature set from the initial group. The former is carried out to reduce the training time, avoid overfitting of data, and finally for simplification of the model. This work is about analyzing the effect of feature selection techniques, Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) on Hierarchical Agglomerative clustering. [3] suggested that PCA is a useful technique for K-Means Clustering. [4] experimentally proved that SVD classifies a document efficiently and in a precise manner. So, in this work, we compare the effectiveness of both dimensionality reduction techniques on clustering. The clustered patent documents are then evaluated using Precision and Recall and F-score.

The remaining sections of the paper are structured as follows: Sect. 2 presents the related work in the fields of document clustering and dimensionality reduction to overcome sparseness of data. Section 3 discusses about the methodology used for analyzing the quality of clusters. Section 4 reports on experimental results carried out as part of this work. Finally, Sect. 5 concludes this paper and discusses the possible enhancements to the current work.

2 Literature Review

Prior art search aims at querying and retrieving the patents to discover any knowledge existing before the filing of the novel patent application. [5] describes the prior art search along with the needs, their challenges, and also about the three distinct levels involved in prior art searching.

Document clustering is an unsupervised machine learning strategy whereby similar documents are grouped based ondistance metrics. Patent document clustering

includes clustering the data taken from various fields of the patent document [6, 7]. Recent trends in technology and the relationship between the clusters or between the query and the clusters can be discovered by analyzing these document clusters. [8] analyzed Chinese patent documents based on the implicit and explicit structure using Self-Organizing Maps. [9] compared K-Means clustering with the K-Medoids and found that K-Means suits well for summarization of documents. The efficiency of K-Means clustering depends on the initial set of seed points [10] and it does not suit well for large dimensional space and a large number of clusters. Even though agglomerative clustering is slower than divisive K-Means, its performance was good. Scatter/Gather Browsing system [11] employs both K-Means and Agglomerative clustering.

Manually assigned IPC codes for the patent documents [12] are treated as clusters, and their performance was compared with a clusterless language model. Smoothing techniques were employed to improve the accuracy of the model. Most of these documents and the cluster quality suffer from dimensionality problem. The performance of the clustering algorithms was affected significantly due to sparseness and high dimensionality of documents [13]. To overcome that, feature selection or dimensionality reduction techniques are employed.

A comparative study on various unsupervised dimensionality reduction techniques for text retrieval was carried out [14–16] regarding complexity, error, and retrieval quality on different document sets. Their results show that a selection technique followed by a transformation method can substantially reduce the computational cost without impacting the cluster performance. [17] studied the effect of dimensionality reduction techniques like SVD and TF-IDF on K-Means Clustering for BBC News and BBC Sports dataset. [18] proposed an iterative feature selection method and compared term contribution process with a variety of feature selection methods. Among them, PCA is the most common method of multivariate statistical analysis.

3 Analyzing Methodology

The methodology of analyzing the patent document clustering was explained briefly in Fig. 1. It includes the following four phases: Patent Document Preprocessor, Feature Extractor, Agglomerative Clusteror, and Cluster Evaluator.

The web contains millions of patent documents derived from various patent databases across the globe. Whenever the user issues a query to the Google Patent search engine, it displays a set of documents. The user retrieves the top set of patent documents from the displayed results. The patent document contains title, abstract, bibliographic information, description, detailed description, and claims. Since the title and abstract of a patent document convey the subject and the core content, this work focuses only on patent document title and abstract. The retrieved documents were stored in a patent database. A sample of the patent database was shown in Fig. 2.

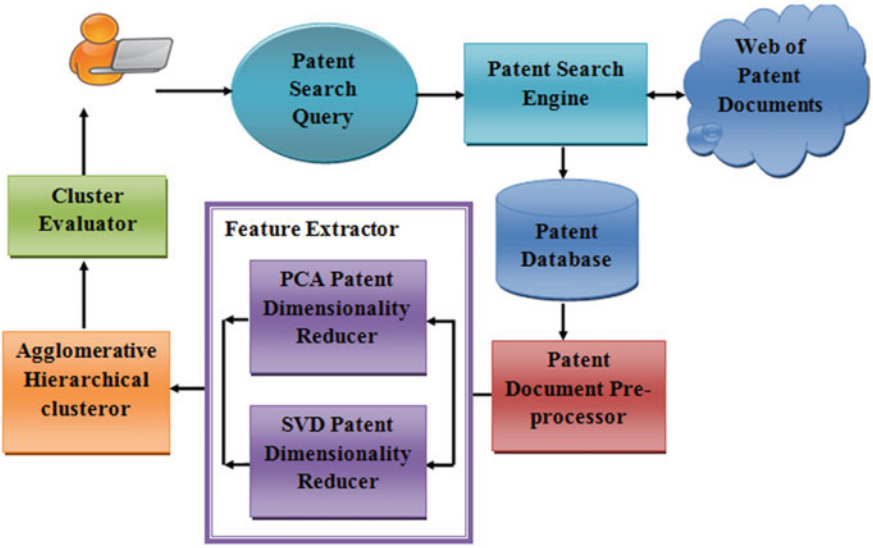


Fig. 1 Methodology of patent document clustering

Patent No	Title	Abstract	IPC
US6728932	Document	Document clustering m	G06F17/(
US2015012	Refining in	A method for computir	G06F17/2
EP1191463	A method	for adapting a k-means	G06K9/62
US9256669	Stochastic	Systems, methods, and	G06F17/3
US7366705	Clustering	Systems and methods f	G06F15/1

Fig. 2 Sample patent database

Preprocessing techniques like tokenization, stopword removal, and stemming [19] are carried out on the document collection in the patent database. This setup makes use of Porter stemmer algorithm [20] for grouping the related forms of a word [11]. The features(words) from the preprocessed patent documents are extracted using PCA dimensionality reducer and SVD dimensionality reducer. PCA dimensionality reducer reduces the dimensionality of the dataset by transforming it to a set of linearly uncorrelated features (words) called principal components. The new set of variables (principal components) will be less than or equal to the initial set, and the principal components retain much information about the document. SVD dimensionality reducer reorders the dimensions of the words along the direction of maximum variation. It decomposes the DTM into a set of singular values and singular orthogonal vectors.

The reduced set of features of the patent document was clustered in a hierarchical manner and the effectiveness of these two dimensionality reduction techniques was compared with precision, recall, and F-score.

4 Experimental Results

This work was conducted with the patent documents retrieved from the Google patent search engine by issuing the query (text + document) clustering. The top 500 documents are retrieved from the Google patent search results and are stored for further processing. The results of each subsequent phase are discussed in the following subsections.

4.1 Dataset

The patent database contains exclusively patent document number, title, abstract, and International patent classification code information. The database was checked for redundancy, and multiple versions of the same document are updated with single latest version. The resultant dataset after removing redundant copies retains 492 documents. The dataset records are preprocessed to remove stop words, numbers, and special characters. In addition to this, some patent document specific words like method, apparatus, embodiment, and system were removed. Porter Stemming algorithm turns all the derivational affixes of a word to a single base word. On transforming these documents to DTM, it was found that the DTM has 99% sparsity. DTM contains 492 documents with 870 unique terms for title field alone. Traditional Vector Space Model was used for document representation. So, each document is considered to be a vector in the term space. The documents are weighted using Term Frequency-Inverse Document Frequency (TF-IDF) score.

4.2 PCA Dimensionality Reducer

When PCA is employed to the DTM, 492 principal components (PC) were obtained. Since the number of documents is less than the number of features/dimensions, it considers documents as features and creates principal components for it.

Table 1 depicts the standard deviation, proportion of variance, and cumulative proportion of variance for first five components. It is clear from Table 1 that the first PC retains much information than the rest of the principal components.

Figure 3 represents the same pictorially. We have considered 40 principal components (>0 variance) and agglomerative clustering was applied to it. For computing, the similarity between the documents, cosine similarity was used to determine the similarity between the documents within the cluster and the linkage between them was average. The sample dendogram of the clusters formed in agglomerative clustering was depicted in Fig. 4.

Table 1 Principal components

Principal components	PC1	PC2	PC3	PC4	PC5
Std. dev	3.9592	2.72293	2.63799	2.55319	2.5356
Variance prop	0.0202	0.00954	0.00896	0.00839	0.0083
Cum. prop of variance	0.0202	0.0297	0.03867	0.04706	0.0553

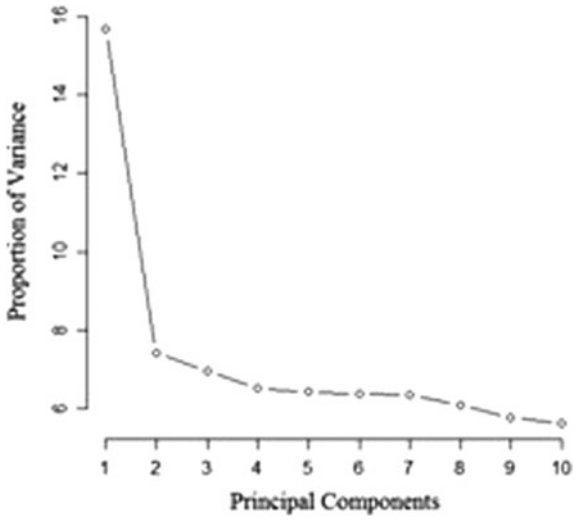


Fig. 3 PCA dimension selection

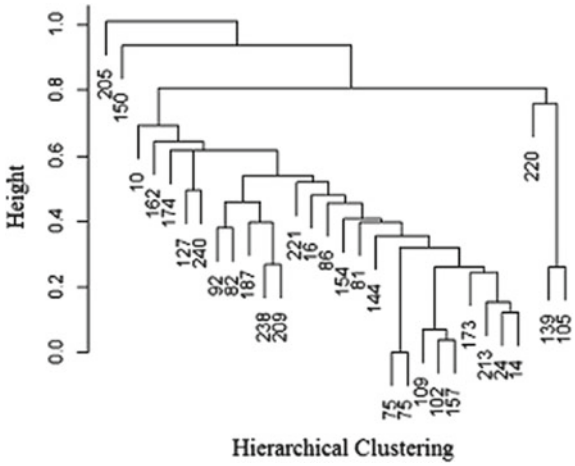


Fig. 4 AHC on PCA

4.3 SVD Dimensionality Reducer

SVD splits the DTM into document-concept similarity matrix, concept matrix, and term-concept similarity matrix. The number of useful dimensions are determined by plotting the results of svd() is shown in Fig. 5. The performance of SVD differs slightly when compared to PCA. The size of the DTM after performing reduction was 492×200. A sample of agglomerative clustered results on reduced dimension set was shown in Fig. 6. This figure depicts that the dissimilarity rate between documents after SVD has reduced to an extent.

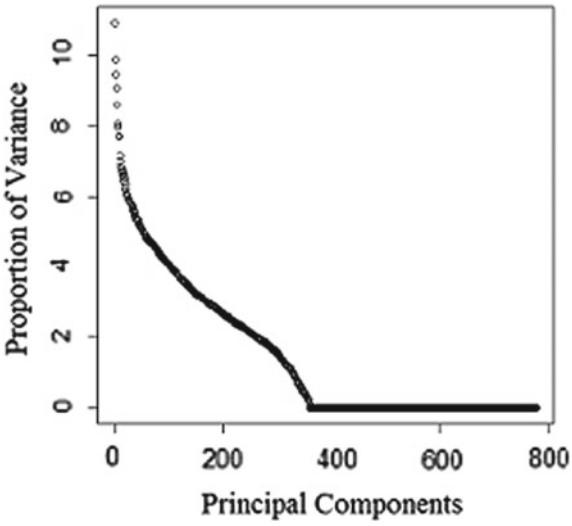


Fig. 5 SVD principal dimensions

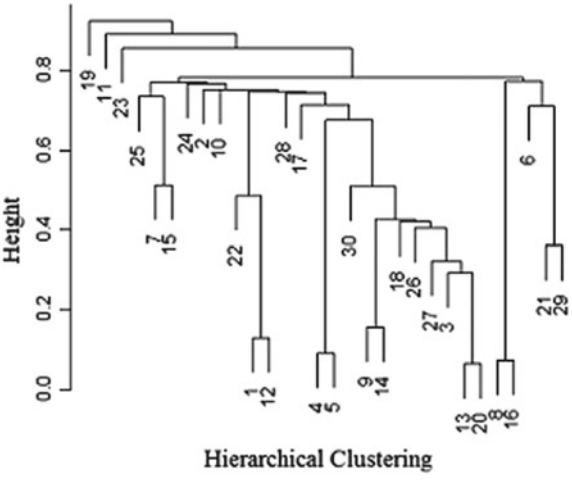


Fig. 6 AHC on SVD

Table 2 Document allocation to clusters

Clusters technique	1	2	3	4
AHC-PCA	426	50	12	4
AHC-SVD	230	108	77	77

Table 3 Cluster evaluation

Cluster measures/technique	Precision	Recall	F-Score
AHC-PCA	0.835	0.7877	0.8116
AHC-SVD	0.829	0.3290	0.8115

4.4 Agglomerative Hierarchical Clusteror

The International patent classification (IPC) code helps to decide the number of clusters. The patent database has four different classes of classification code. Cluster 1(G class) contains a lot of documents because its of type document clustering. Cluster 2 (H class) contains documents related to clustering in network domain, Cluster 3 (A class) denote clustering in medical field while fourth cluster (H class) was about clustering of chemical compounds.

Table 2 shows the grouping of documents into clusters. The majority of the documents (70%) belong to the class G. Proportionately very few documents belong to class A, B, and H. It is clear that from Table 2 of PCA-document grouping 86% documents belong to Cluster 1 because they contain the keywords cluster, document, text, features and so on. Cluster 2 contains documents with terms overlay, server, protocol, network, and so on. Cluster 4 contains four documents not of any specific area. With PCA, there is no even grouping of documents because the terms are not considered as features. But with SVD, the distribution of documents is even and proportionate to that in the patent database. Since there were some misclassifications, the recall of the AHC-SVD is reduced but overall score was good. Table 3 describes the quality of clusters. The clustering results show that the performances of both PCA and SVD are almost similar regarding precision and F-score. The same is visualized from the Figs. 4 and 6. However, PCA is useful when the number of documents is greater than the number of terms.

5 Conclusion

This paper presented an experimental study of two commonly used dimensionality reduction technique PCA and SVD and analyzed the effectiveness of it on

agglomerative hierarchical clustering. The AHC was used because it does not need the number of clusters and its efficiency was better for large datasets in spite of its time complexity. From the results of clustering, it was shown that performance of SVD was better than PCA even though it is expensive. Moreover, PCA is not applicable to highly sparse documents, the number of documents will be less than the number of features(dimensions). In that case, it considers the documents as features and performs the reduction. This work can be extended by analyzing multiple fields of the patent document on a probabilistic language model.

References

1. Mardia, K.V., Kent, J.T., Bibby, J.M.: Multivariate Analysis. Academic Press, London (1979)
2. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann, Elsevier (2011)
3. Ding, C., He, X.: K-means clustering via principal component analysis. In: 21st International Conference on Machine Learning (ICML-04), p. 29. ACM (2004)
4. Li, C.H., Park, S.C.: An efficient document classification model using an improved back propagation neural network and singular value decomposition. Expert Syst. Appl. **36**(2), 3208–3215 (2009)
5. Gaff, B.M., Rubinger, B.: The significance of prior art. Computer. **8**, 9–11 (2014)
6. Jun, S., Park, S.S., Jang, D.S.: Document clustering method using dimension reduction and support vector clustering to overcome sparseness. Expert Syst. Appl. **41**(7), 3204–3212 (2014)
7. Andrews, N.O., Fox, E.A.: Recent developments in document clustering. Technical Report TR-07-35. Department of Computer Science, Polytechnic Institute & State University (2007)
8. Huang, S.H., Ke, H.R., Yang, W.P.: Structure clustering for Chinese patent documents. Expert Syst. Appl. **34**(4), 2290–2297 (2008)
9. Balabantaray, R.C., Sarma, C., Jha, M.: Document clustering using K-Means and K-Medoids. Int. J. Knowl. Based Comput. Syst. **1**(1) (2015)
10. Bradley, P.S., Fayyad, U.M., Reina, C.: Scaling clustering algorithms to large databases. In: 4th International Conference on Knowledge Discovery and Data Mining (KDD-98), pp. 9–15 (1998)
11. Cutting, D.R., Karger, D.R., Pedersen, J.O., Tukey, J.W.: Scatter/Gather: A Cluster-based approach to browsing large document collections. In: 15th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR92), pp. 318–329. ACM (1992)
12. Kang, I.S., Na, S.H., Kim, J., Lee, J.H.: Cluster-based patent retrieval. Inf. Process. Manag. **43**(5), 1173–1182 (2007)
13. Aggarwal, C.C., Yu, P.S.: Finding Generalized Projected Clusters in High Dimensional Spaces. ACM, vol. 29, no. 2 (2000)
14. Mugunthadevi, K., Punitha, S.C., Punithavalli, M.: Survey on feature selection in document clustering. Int. J. Comput. Sci. Eng. **3**(3), 1240–1241 (2011)
15. Kumar, C.A.: Analysis of unsupervised dimensionality reduction techniques. Comput. Sci. Inf. Syst. **6**(2), 217–227 (2009)
16. Tang, B., Shepherd, M., Heywood, M. I., Luo, X.: Comparing dimension reduction techniques for document clustering. Adv. Artif. Intell. 292–296 (2005)
17. Kadhim, A.I., Cheah, Y.N., Ahamed, N.H.: Text document preprocessing and dimension reduction techniques for text document clustering. In: 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology (ICAIET), pp. 69–73. IEEE (2014)

18. Liu, T., Liu, S., Chen, Z., Ma, W.Y.: An evaluation on feature selection for text clustering. ICML **3**, 488–495 (2003)
19. Kantrowitz, M., Mohit, B., Mittal, V.: Stemming and its effects on TFIDF ranking (poster session). In: 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 357–359. ACM (2000)
20. Porter, M.F.: An algorithm for suffix stripping. Progr. Electron. Libr. Inf. Syst. **40**(3), 130–137 (1980)

SC^2: A Selection-Based Consensus Clustering Approach

Arko Banerjee, Bibhudendu Pati and Chhabi Rani Panigrahi

Abstract Consensus clustering, also called clustering ensemble, is a method of improving quality and robustness in clustering by optimally combining an ensemble of clusterings generated in different ways. In this work, we introduce our approach that is based on a selection-based model and use cumulative voting strategy in order to arrive at a consensus . We demonstrate the performance of our proposed method on several benchmark datasets and show empirically that it outperforms some well-known consensus clustering algorithms.

Keywords Clustering · Consensus clustering · Clustering ensemble · Voting

1 Introduction

Consensus clustering is used to improve accuracy and robustness of traditional clustering algorithms. The necessity for consensus clustering arises due to the fact that there exists no universal clustering algorithm that can yield satisfactory clustering result for any given input data. Consensus clustering considers as input an ensemble of clusterings generated by different traditional clustering algorithms and outputs a

A. Banerjee (✉)
College of Engineering and Management, Kolaghat, WB, India
e-mail: arko.banerjee@gmail.com

B. Pati · C. R. Panigrahi
C.V. Raman College of Engineering, Bhubaneswar, Odisha, India
e-mail: patibibudhendu@gmail.com

C. R. Panigrahi
e-mail: panigrahichhab@gmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_18

single robust clustering. Most often the ensemble used as input is overly generated and thus considering the whole ensemble may degrade the quality of output consensus. Naturally, the question arises whether it is necessary to selectively consider some of the clusterings in the ensemble while generating a consensus. This work attempts to solve this problem in a novel way.

The proposed approach is a sort of voting aggregation. Normally, different clusterings differ among themselves and we gain from such diverse opinion. Sometimes, the clusterings are too widely diverse to arrive at a meaningful consensus. Since the objective is to combine clusterings that are qualitatively better than the individual clusterings, it may be worthwhile ignoring some clusterings that are not positively contributing to a meaningful consensus. Our algorithm adopts this principle and unlike the existing voting-based consensus clustering techniques, we selectively and iteratively choose one clustering at a time while ensuring that the internal quality of consensus clustering is monotonically nondecreasing. Though we use internal quality of clustering as the weights assigned to the clusterings, but our method is general enough to handle any criterion function. The usual objection in the voting method of relabeling the clusters is handled by using Hungarian method in our formulation.

The rest of the paper is organized as follows. In Sect. 2, the problem formulation of generic consensus clustering is discussed. Section 3 deals with the proposed technique of selection-based consensus clustering. The experimental results of the proposed method are reported in Sects. 4 and 5 draws the conclusion.

2 Consensus Clustering

Formally, the problem of consensus clusterings can be described as follows: Let S be a set of n data points denoted as $S = \{s_1, s_2, ..., s_n\}$. Given a set P consisting of T clusterings represented as $P = \{P_1, P_2, ..., P_T\}$ of the data points in S. A clustering P on S is defined as $P = \{C_1, C_2, ..., C_k\}$ such that $C_i \subseteq S(\forall i)$, $C_i \cap C_j = \phi(i \neq j, \forall i,j)$ and $\bigcup_{i=1}^{k} C_i = S$. Our goal is to find a final clustering $P^* = \{C_1^*, C_2^*, ..., C_k^*\}$ that optimizes a consensus function. A consensus function maps a P to a final clustering P^*. The P^* is a sort of median of P. The problem of selection-based consensus clustering is formulated in a generic sense as follows: Suppose a set P consisting of T clusterings of the given data points in S. The goal is to find P^* that optimizes a criterion function over the power set of $P(2^P)$. This consensus clustering problem is more general than the earlier formulation and leads to a new area of investigation. In the next section, the proposed method of selection-based consensus technique is described and expressed with the help of the said notations.

3 Selection Based Consensus Clustering

In this section, we introduce a rank matrix for a clustering and is used to define a cumulative voting-based consensus of a set of clusterings. We elaborate our earlier formulation of generic selection-based consensus clustering problem and give a more specific formulation. Given a clustering P on S. From this *hard* clustering, we generate a soft clustering such that μ_{ij} denotes the weight of the data point s_i belonging to the cluster C_j. We assume the weights μ_{ij} are normalized such that $\sum_j \mu_{ij} = 1$. For a clustering P, we define μ_{ij} for each s_i $(1 \leq i \leq n)$ and for each C_j $(1 \leq j \leq k)$ as a metric distance between s_i and the mean of C_j which is being normalized by sum of distances of s_i with all C_js. In the experimental section, the Euclidean distance is used for computing μ_{ij}. The *rank* of C_j for s_i, denoted by $rank(s_i, C_j)$, is the position of C_j when μ_{ij} $(1 \leq j \leq k)$, are arranged in decreasing order. For a clustering P, we construct a $n \times k$ rank matrix U with a row for each s_i $(1 \leq i \leq n)$ and a column for each C_j $(1 \leq j \leq k)$ and each entry u_{ij} is $rank(s_i, C_j)$. For a given clustering P, we can compute the corresponding rank matrix U and similarly, for a given rank matrix U, we can generate the corresponding clustering. The algorithm to find a clustering from a rank matrix is given in Algorithm 1.

Algorithm 1 To find a clustering from a rank matrix

```
INPUT: U
OUTPUT: P = {C_1, C_2, ..., C_k}
C_j = φ, ∀j
for i = 1 : n do
    j̄ = argmax_h U_ih
    C_j̄ = C_j̄ ∪ {s_i}
end for
return P
```

Let $P_1 = \{C_1^1, C_2^1, ..., C_k^1\}$ and $P_2 = \{C_1^2, C_2^2, ..., C_k^2\}$ be two clusterings on S. We assume that the clusters in both clusterings are labeled so that C_i^1 and C_i^2 are corresponding clusterings. This correspondence problem can be solved easily by relabeling the clusters of P_2 with respect to P_1, by applying Hungarian matrix method. Assuming that the clusterings are properly labeled, we define the process of cumulative voting-based consensus. The consensus is obtained through the rank matrix and we first aggregate the rank matrices and use Algorithm 1 to get the consensus clustering from the aggregated rank matrices. Let the rank matrices corresponding to P_1 and P_2 be U^1 and U^2, respectively. The cumulative voting-based consensus of P_1 and P_2, denoted as $P_1 \oplus P_2$, is defined as the clustering obtained from the rank matrix $U^1 + U^2$ using Algorithm 1. The selection-based consensus function maximizes the internal quality of cumulative ranking consensus over the power set 2^P, which is not computationally feasible.

We propose here a greedy algorithm to determine the consensus clustering in polynomial time. Given a set of T clusterings $P = \{P_1, P_2, ..., P_T\}$ of S, we compute

the internal clustering quality (IQ) of each clustering. Evaluating the internal quality of clustering is nontrivial and ill-posed task [1]. Hence, it is difficult to determine an ideal way of evaluating the internal quality of clusterings. In the present study, we consider the internal cluster quality as the ratio of intercluster distances to intracluster distances. Without loss of generality, let $IQ(P_1) \geq IQ(P_2) \geq IQ(P_3) \geq \cdots \geq IQ(P_T)$. In our method, we initialize with the clustering that has the highest internal quality and iteratively perform consensus with clusterings only if the internal quality of the consensus is monotonically nondecreasing. Algorithm 2 takes the corresponding rank matrices of the partitions in P as its input and outputs the rank matrix U of SC^2 which (the clustering) is obtained by using Algorithm 1. The optimal permutation π can be obtained by applying the Hungarian method in cubic time complexity.

Algorithm 2 SC^2: Selection-based Consensus Clustering

INPUT: $\{U^1, U^2, ..., U^T\}$
OUTPUT: U
$U = U^1$, where $U = (u_{ij})_{n \times k}$
for $m = 2 : T$ **do**
  $\pi_m = argmin_\pi \sum_j \sum_i |u_{ij} - u^m_{i\pi(j)}|$ {π represents a permutation of the columns of U}
  $\overline{U} = U + u^m_{i\pi_m(j)}$
  if $IQ(\overline{U}) > IQ(U)$ **then**
    $U = \overline{U}$
  end if
end for
return U

To show that clustering quality does not satisfy monotonicity property, we consider an ensemble of 100 partitions of Iris dataset [2] which is generated with random initializations of k-means algorithm. Out of the 100 partitions, only 3 partitions are found to be distinct. We take the entire ensemble and form the consensus using CSPA [3]. The IQ of the consensus clustering is found to be 0.5021; whereas the quality of the consensus clustering by considering the last two partitions is found to be 0.3826. This leads to the conclusion that clustering quality does not improve if any subset of the ensemble is taken; hence subset of the ensemble should be chosen judiciously.

As an illustration of the goodness of SC^2, we consider Cassini dataset [4] and on it, we compare our method with the well-known Iterative Probabilistic Voting Consensus (IPVC) proposed by Caruana et al. [5]. The mechanism of IPVC is as follows: For each data point s_i in S, a corresponding T-dimensional feature vector is constructed, where the ith feature is simply the cluster label from the clustering P_i. In each iteration of the algorithm, each data point is reassigned to different clusters based on a defined distance function between the considered data point and the previous established clusters in the target consensus clustering. The main operation in each iteration is to find a closest cluster for each data point via a defined distance measure between them. The Cassini data contains 1000 two-dimensional points grouped into three well-seperated clusters but the distribution of data is such that it offers

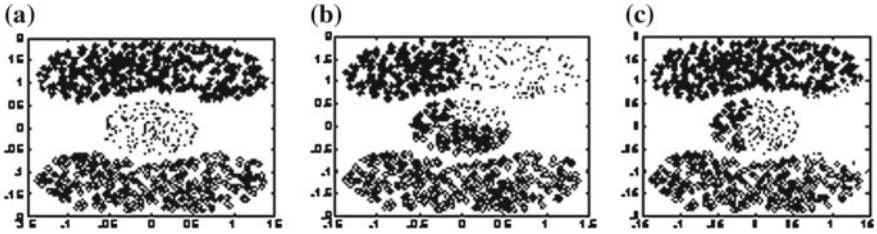


Fig. 1 **a** Three ground truth clusters of Cassini data. **b** Consensus clustering due to IPVC. **c** Consensus clustering due to SC^2

challenging task to any good clustering algorithm. With random initialization of k-means over the data, an ensemble of 100 clusterings is generated. Consensus on the ensemble due to IPVC and SC^2 are shown in Fig. 1 which clearly shows superior performance of SC^2 over IPVC. The Adjusted Rand Index (ARI) [6] on the consensus due to IPVC and SC^2 are found to be 0.512 and 0.7836, respectively.

4 Experiments

The proposed concept of selection-based consensus is not dependent on the method of ensemble generation. To build our ensemble, we used the k-means algorithm as our base algorithm. With different random initialization of k-means on the same data, different clustering solutions are obtained. In this work, we used Adjusted Rand Index (ARI) as an external consensus criteria that measures how well the target clustering performs in comparison to the external true label of the data points. In our experiment, we generate three different ensembles of size 100 for all datasets and compare the average performances of four existing consensus clustering methods CSPA [3], MCLA [3], IPVC [5] and CAS[7]+CSPA with SC^2.

Table 1 shows performances of different consensus algorithms applied on the entire ensemble. In Table 1, CAS+CSPA refers to consensus due to CSPA on a subset selected by Cluster and Select (CAS) proposed by Fern et al. [7]. Fern et al. proposed three subset selection approaches based on external quality and diversity of clusterings. It was shown empirically that among the three methods, CAS method shows the best overall performance. In CAS, similar clusterings are grouped together. The clustering with the highest external quality is selected from each group to form the ensemble. Finally, consensus on the ensemble is done using some known consensus algorithm. But such method suffers from serious disadvantages. The performance of CAS is sensitive to the similarity measure and it is hard to find an optimal similarity measure for a specific dataset. Moreover, the quality of consensus clustering obtained from CAS is dependent on the external consensus algorithm. We observed from experiments (not reported) that some consensus algorithms (which are different from graph-based technique) do not yield satisfying results with CAS. On the

Table 1 Comparison of performances of different consensus clustering algorithms

Data [2]	MCLA	CSPA	IPVC	CSPA+CAS	SCC
Iris	0.7302	0.714	0.7302	0.7148	0.7302
Glass	0.5523	0.552	0.5466	0.5494	**0.5619**
Wine	0.3711	0.3711	0.3711	**0.4437**	0.3711
Chart	0.582	0.5393	0.5263	0.5455	**0.5898**
Ecoli	0.3476	0.2957	0.3981	0.3929	**0.5058**
Segmentation	0.3828	0.3622	0.3877	0.3525	**0.4318**
Yeast	0.1193	0.097	0.1349	0.1009	**0.1649**
Cassini	0.7828	0.6692	0.512	0.7243	**0.7836**

other hand, Strehl and Ghosh [3] defined the cluster ensemble problem as an optimization problem and maximize the normalized mutual information of the consensus clustering. They introduced three different algorithms to achieve good consensus clustering, namely HyperGraph Partitioning (HGPA), Cluster-based Similarity Partitioning (CSPA), and Meta-Clustering (MCLA) algorithms, out of which last two superior methods are considered in our experimentation. The outcome as given in Table 1 indicates that SC^2 outperforms other methods for most of the datasets. Note that CAS reduces the quality of clusterings due to CSPA on glass and Segmentation datasets.

5 Conclusion

In this work, authors proposed a new cumulative voting ensemble approach based upon the ranking of object-cluster associations. Authors also suggested that internal quality being a good clustering measure helps to select significant clusterings that may result a meaningful consensus. The only drawback of the proposed method is due to its cubic time complexity (Hungerian method) for which it may not be scalable for clustering large document datasets.

References

1. Fred, A.L.N., Jain, A.K.: Robust data clustering. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 128–133. USA (2003)
2. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/
3. Strehl, A., Gosh, J.: Cluster ensembles a knowledge reuse framework for combining partitionings. In: Proceedings of AAAI, p. 9398. Edmonton, Canada (2002)
4. Hornik, K.: A CLUE for CLUster ensembles. J. Stat. Softw. **14**(12), 1–25 (2005)
5. Nguyen, N., Caruana, R.: Consensus clusterings. In: Proceedings of the 7th IEEE International Conference on Data Mining, pp. 607–612. Omaha, NE (USA) (2007)

6. Steinley, D.: Properties of the Hurbert-Arabic adjusted rand index. In: Psychol Methods, vol. 9, pp. 386–396 (2004)
7. Fern, X.Z., Lin, W.: Cluster ensemble selection. In: Proceedings of SIAM Data Mining, pp. 787–797. Atlanta, Georgia, USA (2008)

Isolated Kannada Speech Recognition Using HTK—A Detailed Approach

V. Sneha, G. Hardhika, K. Jeeva Priya and Deepa Gupta

Abstract This paper aims to discuss the development of an isolated word recognizer for the Indian language Kannada. The word recognizer is built using HTK, which is based on HMM. The system is trained using triphone HMMs for Kannada words in open space environment from 10 speakers and tested using data from 4 speakers. This paper also gives a comparison of results between MFCC and LPCC techniques for four different test sets.

Keywords Speech to text for Kannada · Hidden markov model (HMM) HTK · MFCC · LPCC · ASR

1 Introduction

Speech is a basic way of communication between human beings. Due to improvement of science and technology, there is a huge need of communication with machines. The way of interaction with machines is done manually which is cumbersome and time taking. This can be solved by making speech as way of interaction between humans and machines by using appropriate tools.

V. Sneha (✉) · G. Hardhika · K. Jeeva Priya
Department of Electronics and Communication Engineering,
Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Bengaluru, India
e-mail: vallapsneha@yahoo.com

G. Hardhika
e-mail: hardhikareddy@gmail.com

K. Jeeva Priya
e-mail: k_jeevapriya@blr.amrita.edu

D. Gupta
Department of Mathematics, Amrita University, Bengaluru, Karnataka, India
e-mail: g_deepa@blr.amrita.edu

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_19

Automatic Speech Recognition (ASR) is a system which takes speech as input, recognizes the speech, and gives corresponding text as an output. Speech recognition systems can be classified into various categories: based on the relevance of the speaker as speaker dependent, speaker independent, based on the manner of speech as isolated speech recognition, connected speech, continuous speech, and spontaneous speech recognition and also based on the size of vocabulary as small vocabulary, which uses less than 100 words, medium vocabulary with a word size of more than 100–1000 words and large vocabulary system, with a word size of more than 1000s of words. Efficient speaker independent system is the ultimate goal to achieve an ideal human to machine interface, but the hurdles faced in the process are the difference in slangs between humans and the noisy environment. Isolated work recognition is a type of ASR in which input has silence or noise between each word pronounced. The widely used recognition methodologies in ASR are word-based recognition system, monophone recognition system, triphone recognition system. The features from speech can be extracted using various techniques, out of which most efficient and popular techniques are MFCC (Mel frequency cepstral coefficients) and LPCC (Linear predictive cepstral coefficients). Research in speech recognition has seen a lot of growth with availability of many common opens source tools such as Julius, Kaldi, CMU-Sphinx, Hidden Markov Model Toolkit (HTK) [1].

In India, there are about 22 regional languages which have been given official status. A good amount of research is going on in the field of speech recognition for various Indian languages. Extensive research is being carried out in speech recognition in the regional Indian languages such as Hindi, Punjabi, Telugu, Tamil, and many others. This paper aims to provide a speech recognition system in Kannada, which is the eighth most spoken language in India and 27th most spoken language in the world. From the author's view, a mere work is done in the field of speech recognition in Kannada language. So, the efforts are put on the language Kannada for speech recognition using HTK as toolkit.

This paper explains how to build a speaker independent isolated Kannada word recognition for digits using MFCC and LPCC feature extraction techniques. Details on how to create a database and complete step by step explanation for creating an ASR using HTK are given in the paper. The remaining part of the paper is organized into eight different sections; Sect. 2 deals with past research work in Indian languages. Section 3 describes about database preparation. Section 4 deals with the proposed method for Kannada speech recognition system. This elaborates the implementation process step by step using HTK toolkit. Section 5 discusses details about the results and gives the comparative analysis of accuracy between MFCC and LPCC techniques for various speech test sets. Section 6 deals with conclusion and future scope.

2 Related Work

In a multilingual country like India, building an ASR system is a complex task. This is due to the fact that there are many languages spoken and there is no common language that all the population knows. The research on building a speech recognition system in Indian languages is going on for many decades. Few of the important research works are emphasized below.

Implementation of a Hindi speech recognition system with a vocabulary size of about 30 words from 8 speakers is discussed in [2]. The parametrization is done using MFCC and speech recognition system for connected words is done using HTK tool in Hindi language for both speaker dependent and independent system and is tested in room environment. The system resulted with an accuracy of 94.3%.

A Hindi ASR system with a vocabulary size of 100 words with MFCC features in the front end and GMM at the backend of ASR is discussed in [3]. Based on triphone-based acoustic modeling, paper [4] compares isolated, connected, and continuous speech recognition in room environment.

The review paper Markov Modeling in Hindi Speech Recognition System: A Review [5] gives explanation on classical and advanced techniques of Markov modeling. A comparative analysis between various well known approaches is done for Hindi. A graphical user interface (GUI) on a java platform is developed for Punjabi Automatic speaker recognition using HTK toolkit in Punjabi speech to text system [6] for connected words. Word recognition rate of 95.8 and 95.4% in classroom and open space environment were achieved by using model proposed in the paper. Concept of hidden Markov models is referred in [7, 8].

Speaker recognition of five district Kannada speaker accents as well as spoken words is explained in Speaker Accent and Isolated Kannada Word Recognition [9]. The model compares baseline Hidden Markov Model (HMM) with Gaussian Mixture Model (GMM). The model is tested for both noisy (little) and noiseless signal using 7056 signals for training and 3528 signals for testing. An average WRR of 95.56% was achieved for known accent speaker, 90.86% for an unknown accent speaker and average recognition of Kannada speaker accents achieved is 82%.

Even though considerable work is done in many Indian languages, not much work is done in Kannada Language. And none of the papers give a detailed explanation on all the procedure using HTK toolkit. So Kannada speech recognition system calls for further research. The method proposed here is a speaker independent triphone based isolated word recognition system along with a comparative analysis of accuracy between MFCC and LPCC, which also includes a self-explanatory step-by-step guide for an ASR system creation for Kannada. The ASR system is created using speech data from 14 speakers, where each speaker spoke 320 words.

3 Database Preparation

Database preparation is the most critical step for creating an ASR. Since there were no freely available speaker independent isolated speech corpus available, the database was created on our own. The speech corpus is created for 10 Kannada words: Ondhu, Eradu, Mooru, Nallakku, Aidhu, Aaru, Elu, Entu, Ombatthu, and Sunya. The speech has been recorded in open space environment using Audacity [10] toolkit at a sampling rate of 16 kHz. The database is created using voices from 14 speakers (8 female + 6 male) using laptop integrated microphone. Out of 14 speakers 10 were used for training and the remaining 4 for testing.

There were four kinds of testing data created, namely: only female voice data set (2 female), only male voice data set (2 male), mixed voice test set (2 female + 2 male), speaker dependent test set (1 male + 1 female speakers who are used for training). Each speaker was given a set of 45 jumbled sentences created using 10 Kannada digits. Every speaker spoke each word for 32 times (Total words = 32 * 10 = 320 words). Totally, voice samples for 4480 words are collected for both training (3200 words) and testing (1280 words). Voice samples are edited to have silence of at least 5 frame lengths between each pronunciation and also to have amplitudes between 0.5 and −0.5. The audacity tool configuration is set to take audio from MONO channel and give output as 16-bit uncompressed WAV format.

4 Proposed Approach

The three main stages of an ASR system are data preparation, training the HMMs, and testing the data. Data preparation includes recording the data, creating transcriptions, creating dictionary and extracting the features of input speech. Training the HMMs step involves creating monophone HMMs, triphone HMMs, and reestimating of HMMs. The testing phase includes recording training data and running the Viterbi algorithm (HVite tool) to recognize the spoken word. The complete block diagram of proposed ASR system creation is shown in Fig 1.

4.1 Module 1: Data Preparation

Step 1: Creating database. In order to build a set of HMMs, a set of speech data files and their associated transcriptions are required. The speech data was collected from 14 speakers using Audacity tool. The HTK reads various formats of input speech files like WAV, HTK, TIMIT, NIST, AIFF, etc., using HWave HTK tool. The transcriptions for each voice samples collected are written in notepad and saved in mlf HTK format.

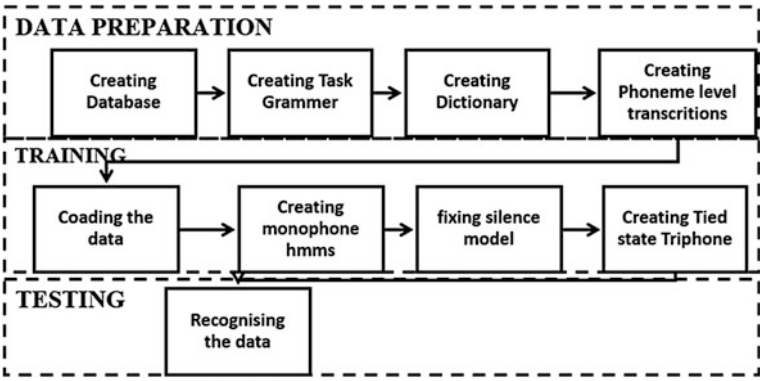


Fig. 1 Block diagram for Kannada ASR creation using HTK

S**tep 2: Creating Grammar**. The grammar for the application is created using notepad. The sentence structure and all the words which for which ASR is trained are given in HTK grammar format. In our application, the sentence structure has a word with silence at start and at end. The words for training are ONDHU, ERADU, MOORU, NALLAKKU, AIDHU, AARU, ELU, ENTU, OMBATTHU, and SUNYA which are the transliterations for English digits 0–9. The grammar file is converted to SLF (standard lattice format) using HParse command for compatibility in HTK.

Step 3: Creating dictionary. The dictionary has words along with their corresponding phonemes in a format accepted by HTK and is created using HLed. The phonemes for all Kannada words are manually created from ARPAbet symbol set, which has a phoneme set of size 39. The word level transcriptions created during the database are converted to phoneme level mlf file using dictionary. During this phone transcription creation, an edit script is given to delete short pause (phoneme-sp) and add silence (phoneme-sil) at start and end of each transcription file. Figure 2 depicts the database collection and dictionary creation.

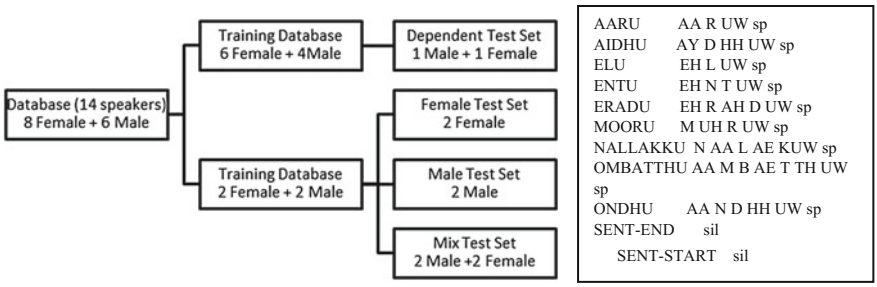


Fig. 2 Kannada database collection and dictionary for Kannada digits

Step 4: Coding the data. The raw speech samples are parameterized into sequences of features. The features are derived using both MFCC and LPCC techniques using HCopy. A total of 39 MFCC features (12-MFCC, 12-ΔMFCC, 12-ΔΔMFCC, P, ΔP, and ΔΔP, where P stands for raw energy of the input speech signal) and 13 LPCC features are derived from each input signal. During the extraction process, a configuration file which specifies the frame length, no of feature, source rate, type of input signal, output signal, etc., are specified.

4.2 Module II: Training Module

Step 1: Creating Monophone HMMs. A prototype model is created for all the monophones of the speech data. Initially, a proto file which contains topology of HMMs (Number of states, means and variances of each state and Transition probabilities) which are to be built is created.

The type of database used is of the type Flat start, i.e., the database does not contain marked sub-word (i.e., phone) boundaries. So all the phone models are initialized to be identical and have state means and variances equal to global speech mean and variance by using HCompV tool. HCompV creates a new proto definition file and vFloors file (Values which set a floor on variances estimated in subsequent steps). From the initial set of models that has been created, the tool HERest when used performs embedded training using the entire training set. HERest tool is given a MMF file and a Macros file as input, MMF file called *hmmdefs* contains a copy for each of the required monophone HMMs. The *hmmdefs* is constructed manually copying the prototype and relabeling it for each monophone (including "sil"). The Macros file contains global options macro and the variance floor macro vFloors generated priorly (The HMM parameter kind and the vector size is given by the global options macro). HERest gives reestimated HMMs for all the phonemes used for training and is done for 4–5 times, to ensure the models are properly trained.

Step 2: Fixing Silence Model. In this step, HMM model for short pause (phoneme-sp) is created by modifying the silence (Phoneme-sil) model. A new sp model is created, which uses center state of sil. A new model for 3-state sp is created in hmmdefs file, by manually making (copy) the central state (state-3) of sil model as center state (state-2) of sp model. Then both the sil and sp models are tied together using HHEd tool. HHEd is file editing tool which applies set of commands given in script to all the HMMs (Fig. 3).

Step 3: Creating Tied State Triphones. A new set of triphones is created from monophone list using HLed tool. And also the word boundaries symbols are mentioned in edit script mktri.led to aid in triphones creation. Then HHed tool is used to tie the monophone together to create HMMs for triphones. So, a new hmmdefs file is created which contains the HMM definitions for all triphone models. Then all the newly created HMMs are reestimated using HERest tool for 4–5 times, to make sure that all the triphone HMMs are properly trained (Figs. 4 and 5).

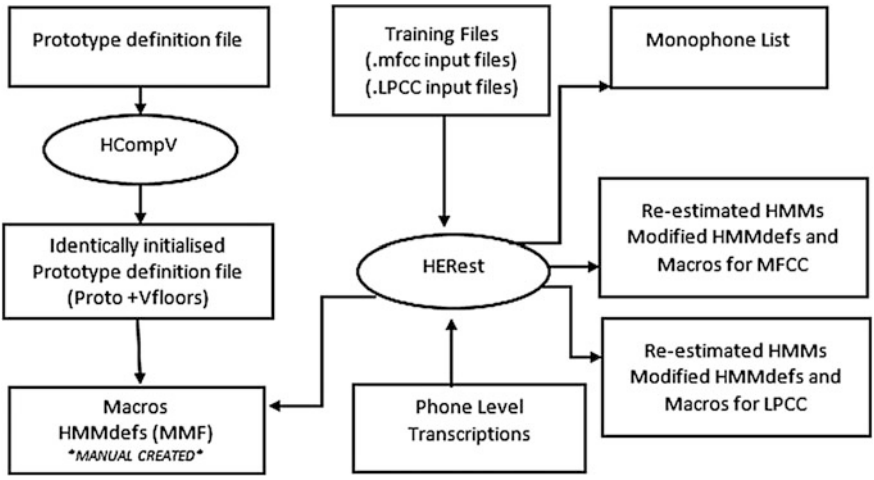


Fig. 3 Block diagram for creating monophone HMMs

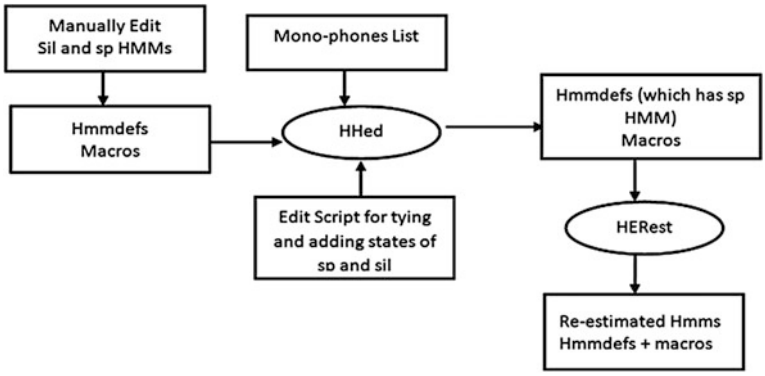


Fig. 4 Block diagram for fixing silence model

4.3 Module III: Recognizing Block

All the paths for feature extracted test files are to be stored in a text file. The recognition is performed using HVite tool. HVite takes a network which describes the allowable word sequences (Wdnet), a dictionary briefing about the pronunciation of each word and a set of HMMs (hmmdef file) as input, performs token passing algorithm in Viterbi Algorithm on the given data and then gives a output mlf file which contains transcriptions of the test data. Then accuracy is calculated manually by comparing the original transcriptions of test data with the recognized transcriptions. Then, analysis of output is done by using accuracy calculated for different test sets (Fig. 6).

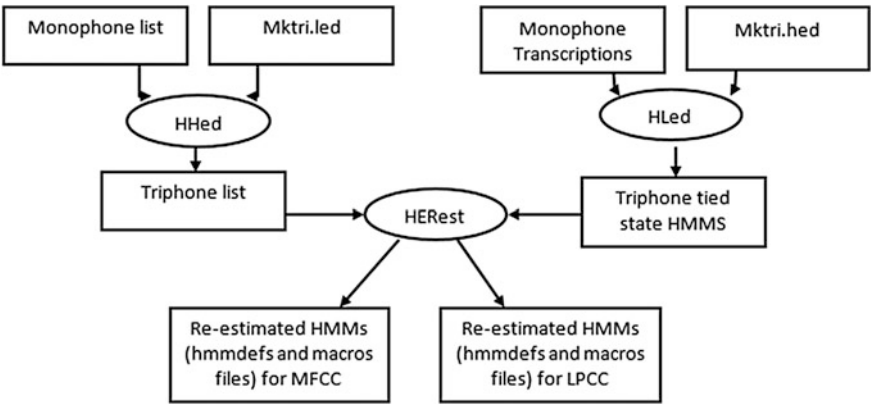


Fig. 5 Block diagram for creating tied state triphones

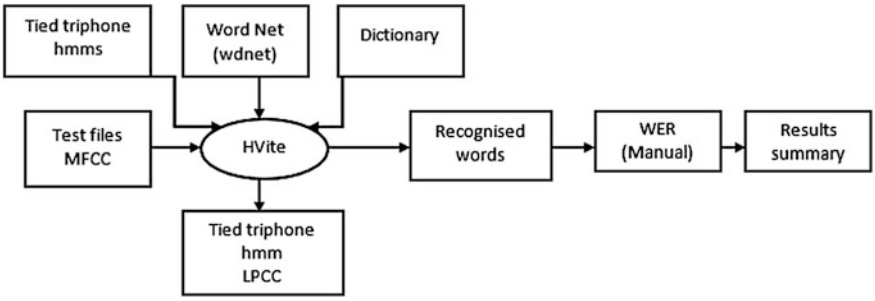


Fig. 6 Block diagram of recognition module

5 Results and Analysis

The system is tested under four kinds of test data, viz., speaker dependent set, where training speech samples are taken from a speaker used in training, mix voices set, which contains test data from both distinct males and females, a male voice set, which contains speech samples from only male speakers, and a female data set, which contains speech samples from only female speakers. The accuracy obtained for speaker dependent case using MFCC is 90% whereas 70% is obtained using LPCC technique. For mixed voice test dataset, the accuracy obtained using MFCC and LPCC techniques are 60% and 30%, respectively (Fig. 7).

For female voice test dataset, the accuracy achieved is 70% for MFCC technique and 40% when LPCC technique is used. And for a male voice data set, the accuracy obtained for MFCC and LPCC technique is 40% and 30%, respectively.

From the analysis graph, we can observe that dependent data set gives very good accuracy when compared to independent test sets. It can also be observed that male

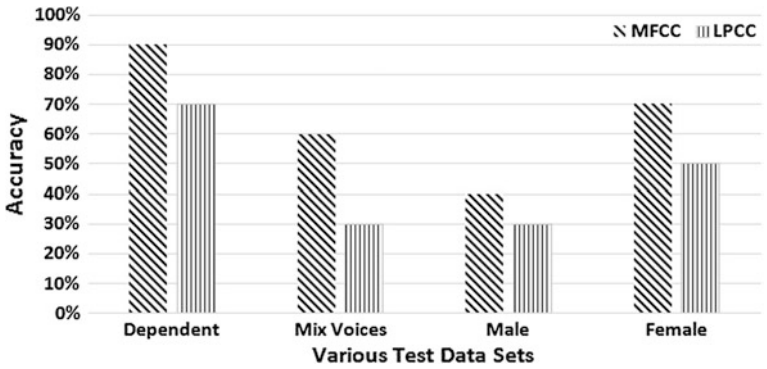


Fig. 7 Comparative analysis of accuracy of speech recognition for different test sets

test set gives a poor accuracy when compared to female test set, this is because the male data collected was from males who have lot of variations in pitch, bass, etc. but the female data collected was taken from speakers having smooth voice (not of high pitch), so in the case of male data collection variation in bass, pitch, etc., between the trained data and the test data resulted in the lower accuracy results. It can be concluded that while creating a database for speaker independent ASR care should be taken in selection of speakers, so that training data has speakers of various kinds of voice.

6 Conclusion and Future Scope

In this paper, a detailed methodology for implementation of the Kannada Speech recognition system for isolated word has been discussed. The accuracy was good when MFCC technique was used when compared with LPCC. The accuracy of the system designed is better for a speaker dependent condition as compared to a speaker independent case. The accuracy for speaker independent case can be improved when better noise reduction techniques are employed for database creation and when more number of states are used for training of phoneme HMMs. It can also be improvised by using Bootstrap data, i.e., data with phone level labeling and also by building decision trees for training. Better results can also be seen when large amount of database covering many slangs is used for training, i.e., when all the multiple pronunciations are included in training database.

7 Declaration

Authors have obtained all ethical approvals from appropriate ethical committee and approval from the subjects involved in this study.

References

1. Hidden Markov Model Toolkit (HTK). http://htk.eng.cam.ac.uk (2012)
2. Tripathy, S., Baranwal, N.: A MFCC based Hindi speech recognition technique using HTK toolkit. In: Proceedings of ICIIP (2013)
3. Kumar, K., Aggarwal, R.K., Jain, A.: A Hindi speech recognition system for connected words using HTK. Int. J. Comput. Syst. Eng. **1**(1), 25–32 (2012)
4. Kumar, A., Dua, M., Choudhary, A.: Implementation and performance evaluation of continuous Hindi speech recognition, ICECS (2014)
5. Aggarwal, R.K., Dave, M.: Markov modelling in Hindi speech recognition system—a review. CSI J. Comput. **1**(1), 34–43 (2012)
6. Dua, M., Aggarwal, R.K., Kadyan, V., Dua, S.: Punjabi speech to text system for connected words. In: Fourth International Conference on Advances in Recent Technologies in Communication and Computing (ARTCom 2012), pp. 206–209. IET
7. Rabiner, L.R.: A Tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE **77**(2), 257–286 (1983)
8. Rabiner, L.R., Huang, B.H.: An introduction to hidden Markov models. IEEE Acoust. Speech Signal Process. Mag. 4–16 (1986)
9. Hema Kumar, G., Punitha, P.: Speaker accent and isolated Kannada word recognition. In: Proceedings of the American Journal of Computer Science and Information Technology. ISSN 2349-3917, 2 Feb 2014
10. http://www.audacityteam.org

Part II
Application of Informatics

Simulation-Based Detection of Lyme Disease in Blood in Rhesus Macaques Using Combined Volterra RLS-MTP Approach for Proper Antibiotic

Sumant Kumar Mohapatra, Sushil Kumar Mahapatra, Santosh Kumar Sahoo, Shubhashree Ray and Smurti Ranjan Dash

Abstract Recently, Rhesus Macaques were infected with Borrelia Burgdor feri (Lyme disease in blood). For detection of residual organisms, various types of methods were utilized including feeding of lab-reared ticks on monkeys, culture, immunofluorescence, etc. To confirm the diagnosis of Lyme disease, usual laboratory test is totally reliable. If diagnosing rate of presence of disease Lyme then there are more possibilities or guidelines are available for treatment. This paper proposed a method for Detecting Borrelia Burgdorferi in Rhesus Macaques by using Volterra RLS Algorithm in addition with multithreading Parallel Approach (MTPA). The proposed method with MATLAB 8.0 is able to accurately detect the Lyme disease in blood with high detection rate. This results in raising certain questions about the pathogenicity of antibiotic-tolerant. From author knowledge, this is the first time to calculate the detection rate of Lyme disease in blood as in infected Rhesus Macaques.

Keywords Tracking · Lyme disease in blood · Volterra RLS algorithm MTPA · Detection rate

S. K. Mohapatra (✉) · S. K. Mahapatra · S. K. Sahoo · S. Ray · S. R. Dash
Trident Academy of Technology, Bhubaneswar, Odisha, India
e-mail: sumsusmeera@gmail.com

S. K. Mahapatra
e-mail: mohapatrasushil@gmail.com

S. K. Sahoo
e-mail: santosh.kr.sahoo@gmail.com

S. Ray
e-mail: shree3269@gmail.com

S. R. Dash
e-mail: sranjandash33@gmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_20

1 Introduction

Sprichetes of the Borrelia Burgdorferi Sensu Lato Species Complex is the root cause of Lyme disease. The early localized, early disseminated, and late stages are the three progression stages in it. Recently, many researchers are investigating the detection and prevention of these three progression stages. Wormser Gary et al. [1] states a simple assessment, treatment, and prevention of Lyme disease. They have issued certain guidelines for Lyme borreliosis therapy. Asch et al. [2] discuss about an infectious and post infectious syndrome. Shadick et al. [3] studied about the population-based retrospective cohort study. Klempner et al. [4] state the two controlled trails of antibiotic treatment in patients symptoms. It also explains about the history of Lyme disease. Philipp et al. [5] state that with early disseminated Lyme Borreliosis C6 antibody successfully treated to patients. Some patients are not affected or they so no change. Fleming et al. [6] state pretreatment and post-treatment assessment of C6 test in patients. Bolz et al. [7] have given a pathway to autoimmunity. Phillip et al. [8] state how Lyme disease infects the human body. Roberts et al. [9] explain how Lyme disease affects the nervous system. Cdavid et al. [10] state about the infection and inflammation on skeletal muscle from nonhuman affected primates. Pachner [11] states about the Lyme meningitis. Cadavid et al. [12] state about the involvement of cardiac in nonhuman primates. Matsui et al. [13] state a comparative pharmacokinesis of ceftazidime, ceftraiaxone, and YM13115 in rhesus monkey also in rats and dogs.

2 Proposed Approach

Our proposed model is based on Volterra RLS Algorithm in addition with MTP Approach. It allows us one thread to run until it executes an instruction that causes latency and then the CPU swaps another threads in while the memory access completes. If a thread does not require a memory access, it will continue to run until time limit is up.

2.1 Mass Center Calculation

The mass center calculation depends on the computation of different moments of the image as:

Initial moment computed by

$$M_{ent}(0,0) = \sum\sum f_{xy} = \sum_{i=0}^{R}\sum_{j=0}^{C}|f_t(i,j) - f_{t+1}(i,j)|, \tag{1}$$

where f_t and f_{t+1} are the tth and $(t+1)$th frame of a video, R and C are number of rows and columns of the frame.

First moment for x and y computed by

$$M_{ent}(0,1) = \sum\sum xf_{xy} = \sum_{i=0}^{R}\sum_{j=1}^{C}|f_t(i,j) - f_{t+1}(i,j)| \tag{2}$$

Second moment for x and y computed by

$$M_{ent}(1,0) = \sum\sum yf_{xy} = \sum_{i=1}^{R}\sum_{j=0}^{C}|f_t(i,j) - f_{t+1}(i,j)| \tag{3}$$

Mean search window location computed by

$$x_{LOC} = \frac{M_{ent}(1,0)}{M_{ent}(0,0)} \tag{4}$$

And

$$y_{LOC} = \frac{M_{ent}(0,1)}{M_{ent}(0,0)}, \tag{5}$$

where x_{LOC} and y_{LOC} are the mean search window locations.

2.2 The Proposed Algorithm

The stepwise implementation of proposed algorithm for detection of Lyme disease in blood (Borrelia Burgdorferi) is as follows:

Step 1: Import multi-moving Lyme disease in blood video from multimedia files.

Step 2: Initialize the population of particles with random values in such a way that they justify the constraints of the control variables as given below

$$W(0) = [000 \ldots 0] \tag{6}$$

$$P(0) = \frac{1}{\delta}\begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix}, \tag{7}$$

where $W(0)$ = Initialize weight

$P(0)$ = Inverse correlation matrix

Step 3: Create n number of thread $T_1, T_2, \ldots T_n$ using Volterra expansion

$$Y(T) = \begin{matrix}[Y(T) & Y(T-1) & \ldots & Y(T-M) \\ Y^2(T) & \ldots & Y^2(T-M) & Y(T) & Y(T-1)]\end{matrix} \tag{8}$$

Step 4: Generate error signal vector using the difference of desired and output signal vector

$$e(T) = d(T) - Y(T)W_K^{T_i}, \tag{9}$$

where each object assign to a specific thread T_i.

Step 5: Inverse correlation matrix can be modified by using MTP approach as given below:

$$\delta(T) = P(T)Y(T) \tag{10}$$

$$P(T+1) = \frac{1}{f_f}\left[P(T) - \frac{\delta_T \delta_T^{T_i}}{f_f + \delta_T^{T_i} Y(T)}\right], \tag{11}$$

where f_f = forgetting factor

Step 6: Estimate the updated vector in two dimension mean window size, i.e.,

$$\widehat{W}(T+1) = \widehat{W}(T) + e(T_1)P(T+1)Y(T) + \ \ldots \ e(T_i)P(T_i+1)Y(T_i) \tag{12}$$

Step 7: Repeat the Step 4 until the value is converged.

3 Result and Discussion

The proposed approach uses the Volterra RLS Algorithm with MTP approach which supports proper diagnosis and measurement of Lyme disease in blood. Table 1 shows the experimentally shown result of proposed method for

Table 1 Output for moving Lyme disease tracking in Rhesus Macaques blood

Input video	Detection rate (D_r)	No. of trails
Lyme disease in blood video	97.28%	0.0928

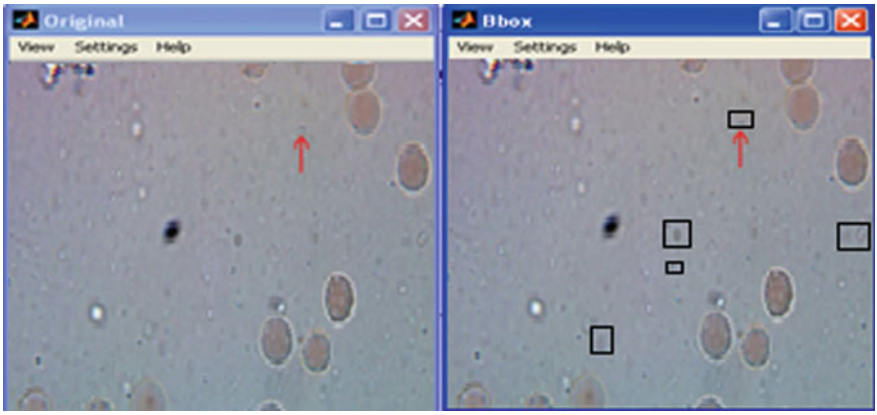


Fig. 1 Simulation result of Lyme disease in blood x100 (frame number 219)

measurement of detection rate on Lyme disease in blood. From this analysis, the given approach has high detection rate of 97.28%. The output for tracking multi moving Lyme disease in blood by proposed approach is shown in Figs. 1, 2, 3, 4, 5, 6, 7 and 8. This paper measures the detection rate for proper diagnosis as:

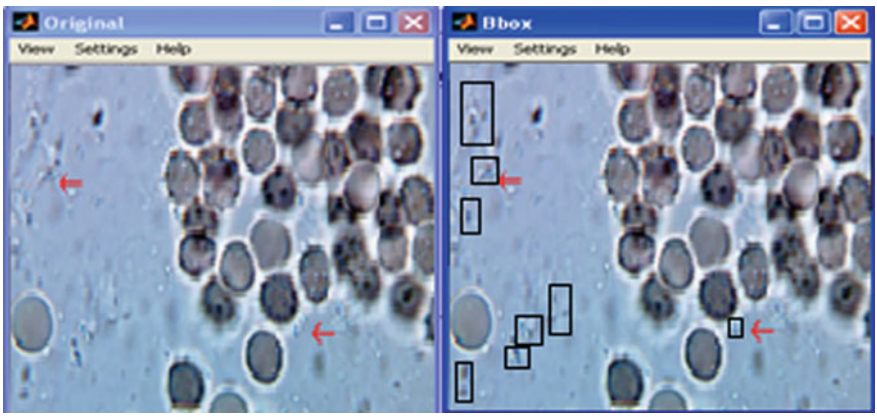


Fig. 2 Simulation result of Lyme disease in blood x100 (frame number 457)

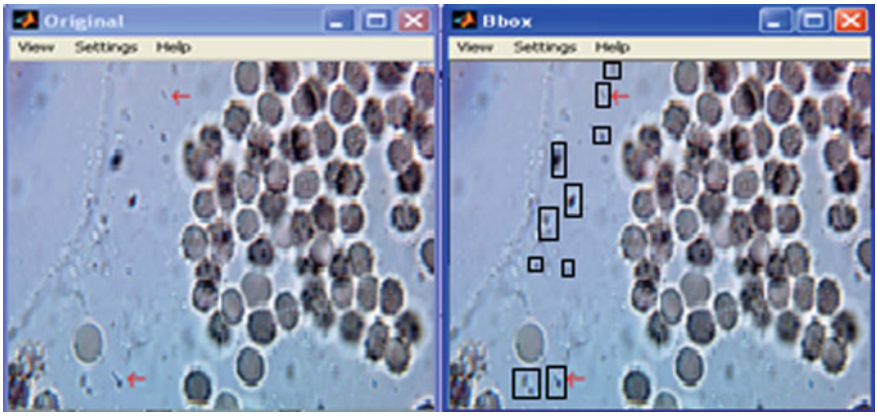


Fig. 3 Simulation result of Lyme disease in blood x100 (frame number 589)

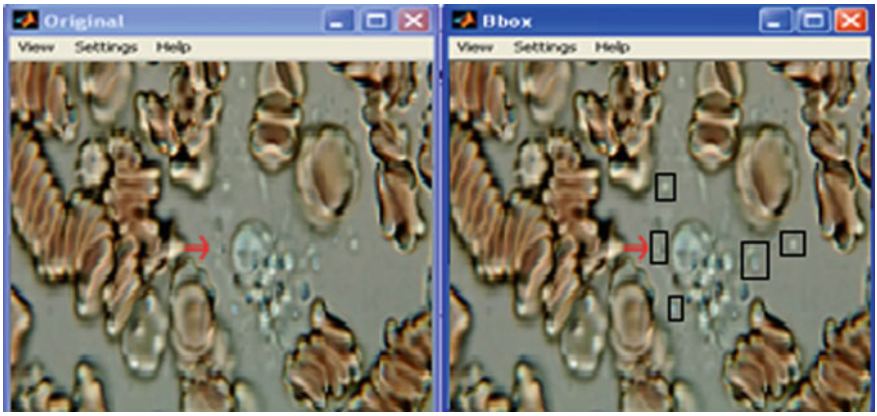


Fig. 4 Simulation result of Lyme disease in blood x100 (frame number 772)

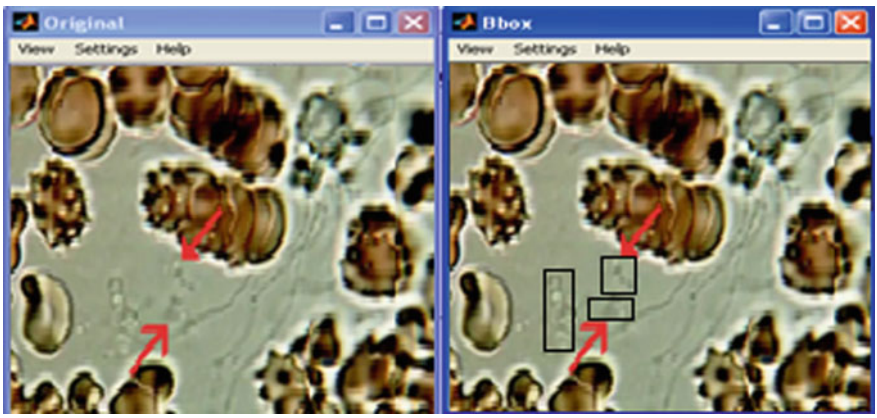


Fig. 5 Simulation result of Lyme disease in blood x100 (frame number 834)

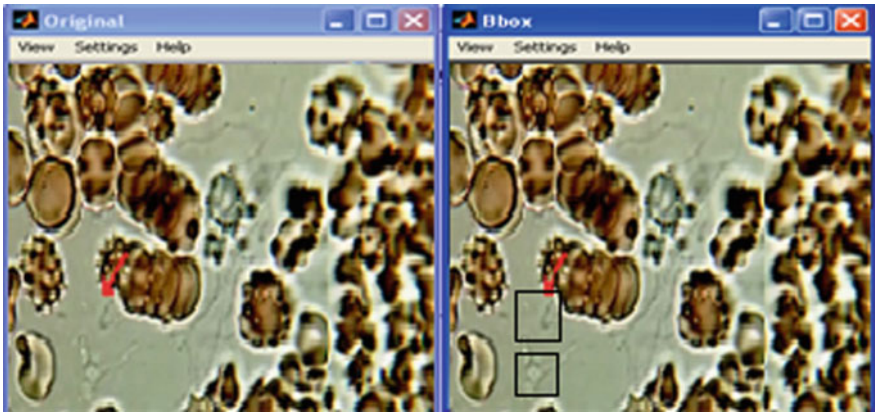


Fig. 6 Simulation result of Lyme disease in blood x100 (frame number 1221)

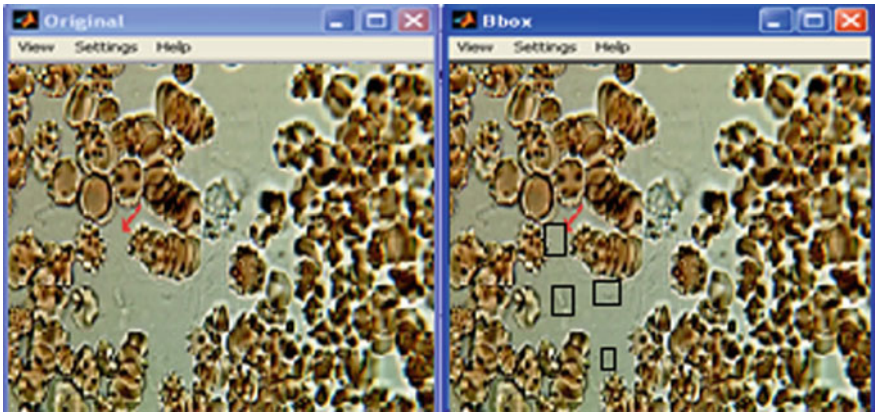


Fig. 7 Simulation result of Lyme disease in blood x100 (frame number 1342)

$$D_r = \frac{T_P}{T_P + T_N}, \tag{13}$$

where D_r = Detection rate of Lyme disease in blood.

T_P = Detected pixel in Lyme disease in blood video.

T_N = Undetected pixel in Lyme disease in blood video.

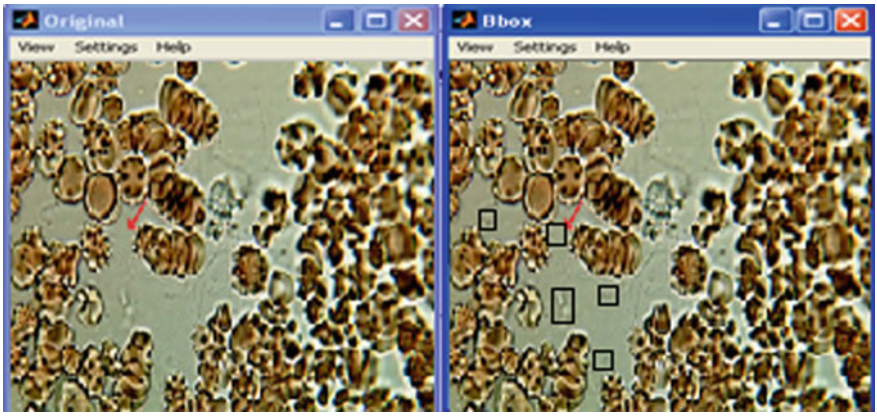


Fig. 8 Simulation result of Lyme disease in blood x100 (frame number 1456)

4 Conclusion

The proposed method is able to track and detect more than one Lyme disease in blood with high detection rate. MATLAB 8.0 results show that the detection result is of more than 97% with Volterra RLS Algorithm with multithreading parallel approach. These results demonstrate that Borrelia burgdorferi can be detected for further antibiotic treatment even if their motion is erratic and fast in blood of Rhesus Macaques. In future, this gives a high response to treatment to overcome the death of forest Rhesus Macaques.

References

1. Wormser Gary, P., Dattwyler Raymond, J., Shapiro Eugene, D., Halperin John, J., Steere, A.: The clinical assessment, treatment, and prevention of lyme disease, human granulocytic Anaplasmosisa and Babesiosis: clinical practice guidelines by the infection diseases society of America. Clin. Infect. Dis. **43**, 1089–1134 (2006)
2. Asch, E.S., Bujak, D.I., Weiss, M., Peterson, M.G., Weinstein, A.: Lyme disease: an infectious and post infectious syndrome. J. Rheumatol. **21**, 454–461 (1994)
3. Shadick, N.A., Phillips, C.B., Logigian, E.L., Steera, A.C., Kaplan, R.F.: The long-term clinical outcomes of Lyme disease. A population-based retrospective cohort study. Annals Internal Med. **121**, 560–567 (1994)
4. Klempner, M.S., Hu, L.T., Evans, J., Schmid, C.H., Johnson, G.M.: Two controlled trials of antibiotic treatment in patients symptoms and a history of Lyme disease. N. Engl. J. Med. **345**, 85–92 (2001)
5. Wormser, G.P., Philipp, M.T., Marques, A.R., Bittkers Marlin, D.S.: A decline in C6 antibody titer occurs in successfully treated patients with culture-confirmed early localized or early disseminated Lyme Borreliosis. Cin. Diagn. Lab. Immunol. **12**, 1069–1074 (2005)

6. Marques, A.R., Fleming, R.V., Klempner, A.R., Schmid, M.S., Dally, L.G.: Pre-treatment and post reatment assessment of the C(6) test in patients with persistent symptoms and a history of Lyme, Borreliosis. Eur. J. Clin. Microbiol. Infect. Dis. **23**, 615–618 (2004)
7. Bolz, D.D., Weis, J.J.: Molecular mimicryto Borrelia Burgdorfer: Pathway to Autoimmunity? Autoimmunity **37**, 387–392 (2004)
8. Aydintug, M.K., Phillip, M.T., Bohm, M.K., Cogswell, R.P.Jr, Dennis, V.A.: Early and early disseminated phases of Lyme disease in the rhesus monkey: a model for infection in humans. Infect. Immun. **61**, 3047–3059 (1993)
9. Bohm, R.P.Jr, Roberts, E.D., Habicht, G., Katona, L.: Pathogenesis of Lyme neuroborreliosis in the rhesus monkey: the early disseminated and chronic phases of disease in the peripheral nervous system. J. Infect. Dis. **178**, 722–732 (1998)
10. Cadavid, D., Bai, Y., Dail, D., Hurd, M., Narayan, K.: Infection and inflammation in skeletal muscle from nonhuman primates infected with different Genospecies of the Lyme disease Spirochete Borrelia Burgdorferi. Infect. Immun. **71,** 7087–7098 (2003)
11. Pachner, A.R.: Early disseminated Lyme disease: Lymemeningitis. Am. J. Med. **98**, 37S–43S (1995) (30S–30S discussion)
12. Bai, Y., Cadavid, D., Hodzic, D., Narayan, E., Barthold, S.W.: Cardiac involvement in non-human primates infected with Lyme disease spirochete Borrelia Burgdorferi. Lab. Invest. **84**, 1439–1450 (2004)
13. Komiya, M., Matsui, H., Ikeda, C., Tachibana, A.: Comparative pharmacokinetics of YM13115, ceftraiaxone, and ceftazidime in rats, dogs and rhesus monkeys. Antimicrob. Agents Chemother. **26**, 204–207 (1984)

A Study on Some Aspects of Biologically Inspired Multi-agent Systems

Gautam Mitra and Susmita Bandyopadhyay

Abstract Nature has always inspired human race in solving various problems. This chapter, in particular, shows a detailed overview of various biologically and nature-inspired multi-agent systems as proposed in the existing relevant literature. A detailed description and analysis of the proposed multi-agent strategies are provided. Multi-agent systems have been in various spheres of management. The applications in terms of cases and situations where these multi-agent systems are applicable are also mentioned along with the introduction of the various agent-based systems. The purpose of this chapter is to assist the readers to think about simulating the existing bio-inspired phenomena with the help of agent-based systems.

Keywords Multi-agent systems · Nature-inspired strategy · Biologically inspired strategy · Nature-based phenomena

1 Introduction

Agent technology is a bio-inspired technology by default since agent was first proposed by imitating the intelligence of intelligent animals including humans. An agent is an intelligent computational system which is long lived, has goals, self-contained, autonomous, and capable of independent decision-making. The characteristics of agents include autonomy, adaptability, cooperation, proactiveness, mobility, social ability, learning ability, veracity, responsiveness, and rationality [1]. Thus agents are supposed to simulate the intelligence level of humans or other intelligent animals. Assigning such characteristics makes the agent

G. Mitra (✉) · S. Bandyopadhyay
Department of Business Administration, The University of Burdwan,
Burdwan 713104, West Bengal, India
e-mail: gautammitra6@gmail.com

S. Bandyopadhyay
e-mail: bandyopadhyaysusmita2010@gmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_21

technology a challenging tool of research in all scientific and technological fields of study. The addition of a facet of intelligence makes a researcher analyze the components of intelligence quotient. Such a difficult task becomes even more difficult because of available technology since implementing such agent-based technology sometimes poses a challenge on the existing technology as well. Since requirement of a technology always depends on the type of the problem under study, thus, the researcher needs to have sound knowledge and firm grip over the required technology, which in turn, imposes challenge on the researchers and practitioners. In many cases, available technology may seem to be insufficient since intelligence levels of intelligent animals including humans have a wide range of various components, such as, interpersonal skills, memory, learning ability, ability to know oneself (intrapersonal ability), logical and analytical ability, spatial ability, ability to deal with novel situations, ability to process information effectively, ability to find solution of a problem, efficiency and precision of one's neurological system [2–4]. Thus, simulating all the components of intelligence quotient in a single agent is really difficult task and extremely challenging. Thus, agent technology is still a growing field of study. The applications of multi-agent systems are observed in various fields of Engineering and Management, as evident from the existing literature. Among the numerous such applications, the works of the researchers in [5–8] can be through of as representative research studies. This chapter focuses its attention on multi-agent systems' simulation only because of its realistic and close analogy to practical problems. Although agent-based technology is based on imitation of the behaviors of intelligent animals, but the choice of such intelligent animal or process may vary. Biology has inspired scientists to device new intelligent machines, intelligent algorithms, and strategies based on various strategies as adopted by various animals in animal kingdom. Even very recently, Williams and Biewener [9] have invented the pigeons' body structure and mechanism which helps them to avoid clash during flight. This strategy may be used to design aircrafts so as to avoid clashes during flight. Various insect-based robots have been made in order to accomplish various tasks imitating the strategies of those insects. A large number of examples can be cited for robot development by imitating the structure of various animals and insects in nature.

A large variety of nature based algorithms and strategies have also been developed as evident from the existing literature, so as to solve various existing problems. An overview of some of the benchmark nature based algorithms as proposed in the existing literature has been summarized by [10]. Some of these algorithms may include genetic algorithm (GA), particle swarm optimization (PSO), artificial immune system (AIS), ant colony optimization (ACO), simulated annealing (SA), honey bee mating algorithm (HBMA), frog leaping algorithm (FLA), firefly algorithm, and so on.

The existing literature has mostly used ant behavior, swarm behavior, immune strategy, honey bee mating strategy, genetic behavior for developing some agent-based systems. Besides these already existing strategies, very few papers have been observed to apply only agent-based technology to simulate any practical biological phenomenon. Thus, the subsequent sections are going to enlighten two

types of bio-inspired agent-based applications—(1) agent-based systems on already existing nature based technology on which various algorithms have been developed, and (2) agent-based technology imitating only existing biological phenomenon.

2 Multi-agent Models Based on Traditional Biological Behaviors

The existing literature shows various algorithms and strategies based on various biological behaviors in the animal kingdom. All of these algorithms and strategies are stochastic in nature in order to simulate the random events occurring in nature. This section explains the applications of various traditional nature-based techniques and their use in developing multi-agent models. The various benchmark nature based algorithms which have been used to build agent based systems are: (1) ant colony optimization (ACO), (2) artificial immune system (AIS), (3) particle swarm optimization (PSO), (4) genetic algorithm (GA), (5) firefly algorithm (FA), and (6) simulated annealing algorithm (SA). Some of these agent-based ideas are provided below.

2.1 Ant Colony Optimization

Ant colony optimization (ACO) is based on the behavior of ants in an ant colony. Ants live in underground place where they live, eat, and mate [10]. The interesting behavior of interest is their foraging behavior for searching food. Ants leave a substance known as "Pheromone" on their path which may be followed by other ants by being attracted by the Pheromone trail. Greater the number of ants on a path, greater is the intensity of Pheromone substance. Since the number of ants traveling on a shorter path per unit time is greater than those traveling on a longer path, thus the ants start to follow the shortest path to food source after a certain period of time. Finding shortest path in this way based on this interesting foraging behavior of ants and division of labor in an ant colony was proposed by [11]. For a detailed understanding of ant colony based algorithm, the work of Dorigo and Stützle [12] may be consulted.

The existing literature shows several research studies simulating ant behavior by agents for various fields of study. Some of the research studies include the studies of Christodoulou [13], Xiang and Lee [14], Gunes et al. [15], and so on. Among these, [13] simulated ACO behavior with multi-agent system and applied the developed model on networking problem. Xiang and Lee [14] simulated ant behavior for coordination and negotiation purpose between various agents such as machine agent, job agent, in a manufacturing situation. Both the machine agent and

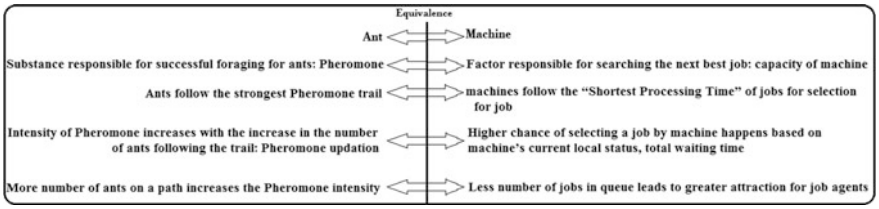


Fig. 1 Analogy of ant behavior with machine agent

the job agent assumed the role of ants. Figure 1 shows the analogy of ant behavior with the machine agent as proposed by [14]. The paper solves task allocation problem in a manufacturing system with the help of machine agents, job agents, and shop agent. Here, the machine agent bids for jobs and task sequencing problem and the job agent bids for processing sequence on the machines. The other papers in the existing literature also show similar application of ant behavior in agent based system.

Figure 1 indicates that the ants in the real world are represented by the machine agents in the agent based system. The factor responsible for searching the next best job for a machine is the capacity of the machine, just like, the substance responsible for successful foraging of ants is called Pheromone. The machines follow shortest processing time (SPT) of the job in order to select a job among several ones, just like, ants follow the strongest Pheromone trail to find out the shortest path to the food source. Here, the Pheromone trail is being compared with the SPT rule. The chance of selecting a job by a machine is dependent on machine's local status and total waiting time, just like, the intensity of Pheromone substance increases with the increase in the number of ants following the trail. Less number of jobs in queue leads to greater attraction for job agents, just like, more number of ants on a path increases the Pheromone intensity. Here, the focus is on number of jobs or ants to a machine or path respectively.

2.2 Artificial Immune System

Immune systems in vertebrates are very complex in nature. Whenever a foreign injurious or toxic agent enters a vertebrate's physical system, then the vertebrate's immunity system protects the body from external attack or infection. The toxic or injurious agent that enters into the physical system is known as antigen whereas the protective agent that is automatically generated to prevent the attack by antigen is known as antibody. This protection mechanism in vertebrates is simulated by artificial immune system where the antigen generally represents inferior solution and antibody represents superior solution. Artificial immune algorithm (AIA) was first proposed by [16] and has been applied in large number of research studies as evident from the existing literature. The algorithm basically selects an antigen

randomly from a population of antigens. Then a sample is taken randomly from a population of antibodies. After this, each antibody in the selected population is matched with the selected antigen and a matching score is computed for the antigen based on the Hamming distance measure. This matching score is added to the fitness value of the antigen. The above process is repeated for a prespecified number of times [10].

The above brief description clearly shows the potential of the immunity system to be implemented by agent-based system, which is realized by several researchers in the existing literature. A significant number of research studies have focused their research work on agent based immune system. Harmer et al. [17] proposed an agent-based defense immune system for the security of computer networks. The authors developed an agent-based prototype in Java in order to implement the proposed approach. Srividhya and Ferat [18] applied agent-based artificial immune system in mine detection and diffusion problem. The authors also developed a solver named AISIMAM (Artificial Immune System based Intelligent Multi-Agent Model) in order to solve the proposed agent-based system. Some of the other significant research studies include the research studies of Dasgupta [19], Castro et al. [20], and so on. The diagram in Fig. 2 shows an example of an analogy of the actual immune system with agent-based artificial immune system as proposed by [17].

Figure 2 indicates that both the antigens and antibodies are represented by agents. Real antibodies are shaped like strings. Similarly antibodies in agent-based system are simply bytes of strings. Agent-based system detects and replaces inferior solution (representing the antigens) by the superior solutions (representing the antibodies), just like, the biological immune system (BIS) detects and destroys antigens or harmful foreign invaders to the physical systems. Immunity-based agent system becomes active whenever there is a virus attack to the computer system, just like, the antibodies become active in case of attack of the external injurious agents into the physical system. Agent system compares antibodies and antigens by string matching rules, just like, BIS matches antigens with antibodies through physical

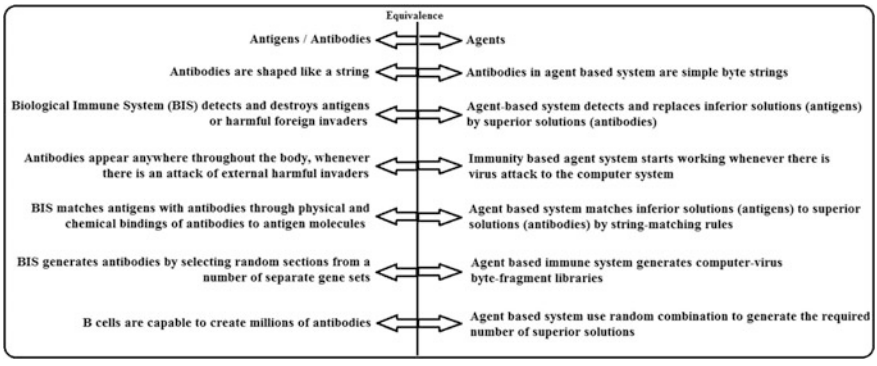


Fig. 2 Analogy of biological immune system with immune-based agent system

and chemical bindings of antibodies to antigen molecules. Agent-based system generates computer virus byte-fragment libraries, just like, BIS generates antibodies by selecting random sections from a number of separate gene sets. Agent-based system uses random combination to generate the required number of superior solutions, just like, in BIS, the B cells create millions of antibodies.

2.3 *Particle Swarm Optimization*

The word "swarm" basically indicates social insects such as, ants, bees, termites, and so on. All of them live in colonies. The behaviors of these insects that are of interest to the researcher are their foraging behavior for searching food, communicating process to communicate with the other insects, networking among the insects and their colonies, task allocation among the insects in a colony, division of labor, transportation of food and other required materials, cooperation among the insects in a colony, their mating behavior, nest building techniques, and so on [10]. Nature has taught us of various techniques and nature has inspired to develop algorithms and strategies to solve various problems of mankind.

Particle swarm optimization (PSO) is an algorithm which simulates the behavior of swarms and was proposed by [21]. The "swarm" basically indicates the "population of solutions" whereas the word "particle" represents "a particular solution". This is a kind of evolutionary algorithm where the solution improves over a number of generations towards the final Pareto optimal solutions [22]. The position of the particles (solutions) is updated (that is, the solution is improved) by modifying the velocity of the particles. Swarm behavior has motivated the researchers all over the world to view this behavior from agent-based system's point of view. Thus, a number of researchers have simulated swarm behavior through agents, as evident from the existing literature. For example, [23] proposed a multi-agent based particle swarm optimization technique and applied the proposed technique on a power dispatch problem. The authors simulated the behavior of bee in this paper. Some of the other significant research studies include the research studies of Selvi and Umarani [24], Minar et al. [25], and so on. The analogy of swarm behavior with the agent-based system as proposed by [23], as an example, is shown in Fig. 3.

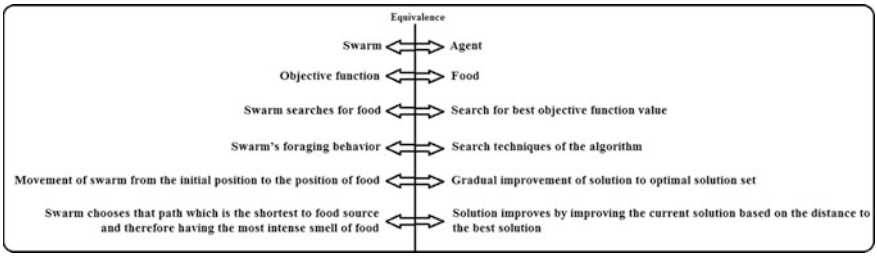


Fig. 3 Analogy of swarm foraging behavior with swarm-based agent system

Figure 3 indicates that the real swarms are represented by agents in an agent-based system. The objective function in the agent-based application represents the food source. Naturally, agent-based system searches for the best objective values, just like, swarms search for food. Swarm's foraging behavior for searching food is simulated by search algorithm. Inferior solutions in agent-based system gradually improve towards the optimality, just like, swarms gradually move to the food source. Solution for the agent-based system improves based on their distance to the best solution, just like, swarm chooses the shortest path towards food source over time.

Besides the above, there are a few other agent-based bio-inspired systems as proposed in the existing literature. Among these, [26] developed a multi-objective genetic algorithm model within a multi-agent framework. Although this research study is not exactly a simulation of genetic theory by agent-based technique, but this study is rather a mix of the agent based system with genetic algorithm. Eguchi et al. [27] simulated symbiotic evolutionary concept with the help of agent-based system.

However, the number of research studies simulating the existing nature-inspired strategy is very few as evident from the existing literature. The reason for such a short number of research studies in this area seems to be the difficulties faced while simulating such system. Such systems are difficult to imagine as well as difficult to implement since in-depth knowledge about various software packages are required for implementation.

3 Multi-agent Models Imitating Existing Natural Behaviors

The existing literature shows almost rare publications simulating an already nonexistent animal behavior. Throughout the entire literature such nature-based strategy which is worth mentioning is the Tarantula mating based multi-agent strategy as proposed by [28]. This paper has also used PROMETHEE multi-criteria decision analysis technique for final decision-making purpose. The paper has proposed a hierarchical multi-agent based framework for simulating Tarantula mating behavior and applied the proposed strategy on three manufacturing networks. The basic idea of this strategy is described below in brief.

In Tarantula mating, the female spider sometimes eats the male spider just after the mating, for genetic purpose or for lack of food. Such strange but interesting behavior has been applied to routing of jobs in a network. Instead of selecting the entire optimum source-to-destination path in a busy network, it is wiser to select the optimum neighbor. This is because of the fact that, in busy network, the entire optimum source-to-destination path may not stay optimum over time because of various factors such as, congestion, possible deadlock condition, buffer status, width of the path, number of intermediate nodes or stations, and so on. In order to

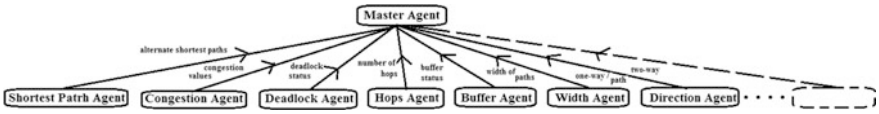


Fig. 4 Hierarchical agent-based framework

simulate the Tarantula mating, the authors have proposed a hierarchical multi-agent framework where the master at the top of the hierarchy takes the final decision by PROMETHEE multi-criteria decision analysis technique on the basis of the various data on the network as delivered by various worker agents at the leaf level of the hierarchy. A generalized hierarchy of such agent-based system is shown in Fig. 4. The routing strategy considered in this research study finds the next optimum neighboring node through agent-based technique, instead of finding the entire source-to-destination path. Thus the worker agents and then the master agent are active and will start functioning whenever there will be a need to route a job to the next optimum neighboring node and whenever a new job enters the system.

All these agents work in parallel. Assuming that a job is waiting on a particular node to be routed from, the shortest path agent calculates the shortest path lengths between each of the immediate neighbors and the current node; congestion agent funds the congestion (represented by the number of jobs) on the arc between the current node and each of the immediate neighbors and between each of the immediate neighbors and the neighbors of the immediate neighbors; the deadlock agent checks whether any of the immediate neighbors is a part of any cyclic deadlock; hops agent counts the number of intermediate nodes between each of the immediate neighbors and the destination; buffer agent checks the status of the buffer at each of machine centers; width agent records the widths of the arcs between the current node and each of the immediate neighbors; direction agent indicates whether any of the arc between the current and the immediate neighbors is one-way or two-way. Similarly one or more other agents can also be added to the above hierarchy which is denoted by the dotted oval and the dotted arrow. Here, the worker agents are shortest path agent, congestion agent, deadlock agent, hops agent, buffer agent, width agent, direction agent, and so on.

The master agent receives all the above data from the worker agents in parallel. After receiving the all the data from the worker agents, the master agent kills them in order to save valuable computational resources. The final decision for selecting the optimum immediate neighbor is accomplished by applying PROMETHEE multi-criteria decision analysis technique. The analogy of the Tarantula mating behavior with the proposed hierarchical multi-agent based system is shown in Fig. 5. At first, the master agent chooses the worker agents, just like, the female Tarantula spider chooses the male partner. After accomplishing the tasks, the worker agents return the required data to the master agent, just like, the male spider transfers the genetic material (sperm) to the female spider. Then the master agent kills the worker agents, just like, the female spider eats the male partner just after mating. The master agent then receives the notification of the death of the worker

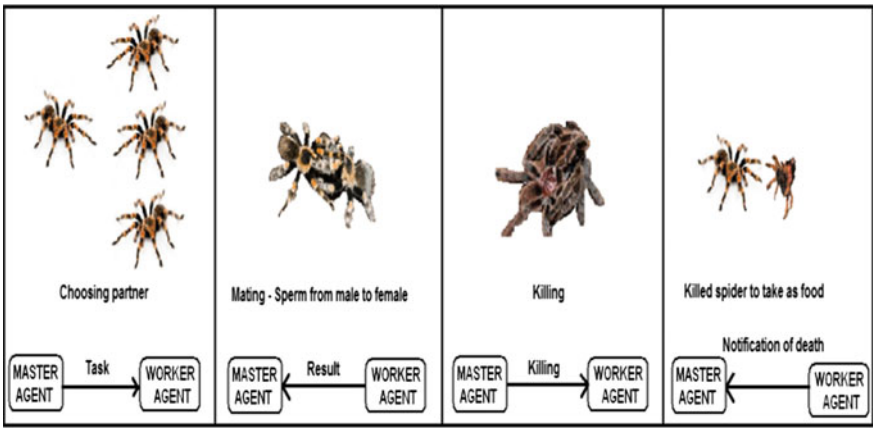


Fig. 5 Analogy with Tarantula mating behavior

agents, just like, the male partner is taken as a food sometimes, by the female tarantula. The above Tarantula mating based strategy may also be applied in various other networks and purposes such as, medical diagnosis, supply chain network, geographical information system, computer networks, supplier selection problems, choice of marketing strategies, choosing financial options, project selection problem, among many others.

4 Glimpse of More Ideas on Nature Based Phenomena

The introduction of the main concepts in the above sections brings the relevance of introducing some new ideas to the future researchers in the respective fields of study. Although the strategies or algorithms on nature-inspired phenomena are abundant in the existing literature, but the application of agent-based technology or the simulation of the existing natural concepts by agent-based techniques makes the implementation more realistic and applicable to the realistic problems. Therefore, this section enlightens some nature-based concepts which may be simulated and applied to the real-world problem solving. There are significant number of books and Internet materials available on hand nowadays and the intense study of these materials can bring new ideas which can be applied to solve practical problems. Some of such references include the books by Vincent and Ring [29], Roots [30], Wilsdon [31], and so on. However, one of the ideas which has attracted the author's attention is delineated below in brief. A particular rat known as desert mole rat shows very strange behavior. In time of danger, they have the ability to run fast forward as well as backward with the same speed blindly. Besides, in order to save the only queen rat which alone can breed, the other female rats are pushed forward to danger and stress. This particular strategy can be used as a security provision in

various computer and other systems, where, several similar dummy modules can be sent along with the main module over the communication network in order to confuse the malicious programs or hackers which is similar to the protection strategy to protect the queen mole. Similar behaviors can be observed if the existing materials are studied thoroughly.

5 Conclusion

This chapter has introduced a variety of agent-based models as proposed in the existing literature. Some of these models are the simulation models of already existing nature based techniques such as ant behavior based model, particle swarm based model, immunity based model, and so on. Besides, Tarantula mating based agent system has also been introduced in this chapter. The reader of this chapter is expected to get an overall idea of how to think about simulating a natural phenomenon by agents, in order to solve practical problems, since this is purpose of this chapter. The chapter concludes with an expectation of the probable future breakthrough ideas on agent-based simulation techniques.

References

1. Bandyopadhyay, S., Bhattacharya, R.: Discrete and Continuous Simulation: Theory and Practice. CRC Press, Florida (2014)
2. Gardner, H., Hatch, T.: Educational implications of the theory of multiple intelligences. Educ. Res. **18**(8), 4–10 (1989)
3. Sternberg, R.: The Nature of Creativity: Contemporary Psychological Perspectives. Cambridge University Press (1988)
4. Perkins, D.: Outsmarting IQ: The Emerging Science of Learnable Intelligence. The Free Press, New York (1995)
5. García-Flores, R., Wang, X.Z.: A multi-agent system for chemical supply chain simulation and management support. OR Spectr. **24**(3), 343–370 (2002)
6. Giannakis, M., Louis, M.: A multi-agent based framework for supply chain risk management. J. Purchasing Supply Manag. **17**(1), 23–31 (2011)
7. Lou, P., Zhou, Z.-D., Chen, Y.-P., Ai, W.: Study on multi-agent-based agile supply chain management. Int. J. Adv. Manuf. Technol. **23**(3–4), 197–203 (2004)
8. Lee, J.-H., Kim, C.-O.: Multi-agent systems applications in manufacturing systems and supply chain management: a review paper. Int. J. Prod. Res. **46**(1), 233–265 (2008)
9. Williams, C.D., Biewener, A.A.: Pigeons trade efficiency for stability in response to level of challenge during confined flight. Proc. Natl. Acad. Sci. **112**(11), 3392–3396 (2015)
10. Bandyopadhyay, S., Bhattacharya, R.: On some aspects of nature-based algorithms to solve multi-objective problems. In: Yang, X.-S. (ed.) Artificial Intelligence, Evolutionary Computation and Metaheuristics. Studies in Computational Intelligence, vol. 427, pp. 477–524. Springer (2013). ISSN: 1860-949X, ISBN: 978-3-642-29693-2
11. Dorigo, M., Gambardella, L.M.: Ant colony system: a cooperative learning approach to the traveling salesman problem. IEEE Trans. Evol. Comput. **1**(1), 53–66 (1997)
12. Dorigo, M., Stützle, T.: Ant Colony Optimization. MIT Press, USA (2004)

13. Christodoulou, S.: Construction imitating ants: resource-unconstrained scheduling with artificial ants. Autom. Constr. **18**(3), 285–293 (2009)
14. Xiang, W., Lee, H.P.: Ant colony intelligence in multi-agent dynamic manufacturing scheduling. Eng. Appl. Artif. Intell. **21**(1), 73–85 (2008)
15. Gunes, M., Sorges, U., Bouazizi, I.: ARA-the ant-colony based routing algorithm for MANETs. In: Proceedings of IEEE International Conference on Parallel Processing Workshops, pp. 79–85. Vancouver, B.C., Canada, 21 Aug 2002 (2002)
16. Bersini, H., Varela, F.J.: A variant of evolution strategies for vector optimization. In: Schwefel, H.-P., Männer, R. (eds.) PPSN 1990. LNCS, vol. 496, pp. 193–197. Springer, Heidelberg (1991)
17. Harmer, P.K., Williams, P.D., Gunsch, G.H., Lamont, G.B.: An artificial immune system architecture for computer security applications. IEEE Trans. Evol. Comput. **6**(3), 252–280 (2002)
18. Sathyanath, S., Sahin, F.: AISIMAM—An Artificial Immune System Based Intelligent Multi Agent Model and Its Application to a Mine Detection Problem. http://scholarworks.rit.edu/other/455 (2002)
19. Dasgupta, D.: Immunity-based intrusion detection system: a general framework. In: Proceedings of the 22nd National Information Systems Security Conference, vol. 1, pp. 147–160. Hyatt Regency Hotel, Crystal City, VA, USA, 18 Oct 1999 (1999)
20. Castro, D., Leandro, N., Jon, T.: Artificial immune systems: a novel paradigm to pattern recognition. Artif. Neural Networks Pattern Recogn. **1**, 67–84 (2002)
21. Eberhart, R.C., Kenndy, J.: A new optimizer using particle swarm theory. In: Proceedings of the Sixth Symposium on Micro Machine and Human Science, pp. 39–43. IEEE Service Center, Piscataway (1995)
22. Coello, C.A.C., Lamont, G.B., Veldhuizen, D.A.V.: Evolutionary Algorithms for Solving Multi-Objective Problems, 2nd edn. Springer, USA (2007)
23. Kumar, R., Devendra, S., Abhinav, S.: A hybrid multi-agent based particle swarm optimization algorithm for economic power dispatch. Int. J. Electr. Power Energy Syst. **33**(1), 115–123 (2011)
24. Selvi, V., Umarani, R.: Comparative analysis of ant colony and particle swarm optimization techniques. Int. J. Comput. Appl. **5**(4), 1–6 (2010)
25. Minar, N., Burkhart, R., Langton, C., Askenazi, M.: The swarm simulation system: a toolkit for building multi-agent simulations. Working Paper 96-06-042, Santa Fe Institute, Santa Fe (1996)
26. Cardon, A., Galinho, T., Vacher, J.-P.: Genetic algorithms using multi-objectives in a multi-agent system. Rob. Auton. Syst. **33**, 179–190 (2000)
27. Eguchi, T., Hirasawa, K., Hu, J., Ota, N.: A study of evolutionary multiagent models based on symbiosis. IEEE Trans. Syst. Man Cybern. Part B: Cybern. **36**(1), 179–193 (2006)
28. Bandyopadhyay, S., Bhattacharya, R.: Finding optimum neighbor for routing based on multi-criteria, multi-agent and fuzzy approach. J. Intell. Manuf. **26**(1), 25–42 (2015)
29. Vincent, H.R., Ring, T.C.: Encyclopedia of Insects. Academic Press, USA (2003)
30. Roots, C.: Nocturnal Animals. Greenwood Press, London (2006)
31. Wilsdon, C.: Animal Behavior: Animal Defenses. Chelsea House Publishers, New York (2009)

A Qualitative Hemodynamic Analysis on Human Cerebrovascular Phantom

Pranati Rakshit, Nirmal Das, Mita Nasipuri and Subhadip Basu

Abstract Patient specific hemodynamic analysis is one of the major factors to determine one's cerebrovascular health. Irregularities in hemodynamic pressure, velocity, and wall shear stress may lead to several cerebrovascular diseases. In this present work, we present a detailed hemodynamic analysis on two cerebrovascular phantoms, an anterior communicating artery, and a complete arterial tree including the *circle of Willis*. We have shown the step-by-step procedure involved in converting the surface mesh to a solid, which is needed for flow analysis using ANSYS Fluent. Then, we have analyzed the hemodynamic parameters like velocity and wall shear stress through flow analysis using various computational fluid dynamics techniques. The developed method is useful for future research on hemodynamic analysis using 3-D digital flows.

Keywords Hemodynamic analysis · Computational fluid dynamics
Cerebrovasculature · Digital flows

1 Introduction

Hemodynamic parameters are alleged to be the leading factors in respect to the genesis and progression of vascular diseases. It is important to understand the hemodynamic factors like velocity, wall shear stress (wss), and pressure that play enormous role in the formation and development of cerebral diseases like aneurysms [1], atherosclerosis, etc. So finding out the vulnerable region which may lead to some anomaly in the health of vascular structure is an important research problem.

The conventional approach to study the hemodynamic parameters is numerical simulation using finite element method [2]. One new approach to analyze the same

P. Rakshit (✉) · N. Das · M. Nasipuri · S. Basu
Department of Computer Science and Engineering, Jadavpur University,
Kolkata 700032, India
e-mail: pranatirakshit@yahoo.co.in

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_22

is using digital flow based model. The work presented in this paper is the first step towards a complete cerebrovascular analysis using the 3-D digital flow based method [3]. We present here the challenges involved in reconstructing the 3-D vascular structure and subsequent processing for flow analysis through finite element methods. More specifically, the major arterial structures present in the human carotid region are focused and several synthetic phantoms of human cerebrovasculature are designed. For the construction of synthetic vascular phantoms, we have adopted Bezier curve based approach as proposed in [4, 5].

To do the hemodynamic analysis, it is important to identify the carotid blood flow characteristics and so therefore the structural analysis of carotid arteries is needed to determine the unwanted and abnormal outpouchings, blockage, or shrinkage, etc. To understand carotid arteries, we need to be familiar with the formation of the *circle of Willis* which is there at the base of the human brain. This is a circular vasculature formed by the right and left *Internal Carotid Artery* (ICA), right and left *Anterior Cerebral Artery*, *Anterior Communicating Artery* (ACA), left and right *Posterior Cerebral Artery*, *Posterior Communicating Artery* (left and right). The *Middle Cerebral Arteries* and the *Basilar Artery* are also considered part of the circle.

For carotid arterial system, in vivo or ex vivo analysis is very complex and time consuming task and it is required to be handled carefully by the expert clinicians. To capture the in vivo images, patient permission and involvement is very necessary and equally important. So as much it is important to analyze real medical data, it is also important to construct synthetic structures or phantoms. Therefore, generation of the synthetic structures or phantoms is often beneficial for the patients who can be escaped from repeated harmful cerebral scans. Phantom use can be extended to experimental evaluation of new computational techniques at the laboratories. Phantom-based simulation experiments are attracting momentum in experimental biology due to its apparent design simplicity. There are two aspects of the study in general: (1) digital modeling of the vasculature, using mathematical models, (2) using physical vascular phantoms, generated by casting a replica of the actual vasculature. Researchers often adopt both the studies to validate their experiments. The previous option is adopted in the current experiment.

In first step, original CT image is taken as input and converted into fuzzy distance transform (FDT) image. One can find a vast and rich literature on FDT and its application [6–11]. Recently, Guha et al. proposed a FDT-based geodesic path propagation algorithm for reconstruction of cerebrovascular phantom which required least user interaction [12]. In next step, start and end input seed points are taken from user and geodesic path is found between those two points. Next sphere is drawn at each point on the geodesic path with radius equal to the FDT value of that point. If generated phantom needs further modification, user can give more seed points and followed same steps described before.

Hemodynamics is actually the fluid dynamics of blood flow. It explains the physical laws that govern the flow of blood in the blood vessels. Due to viscosity, flow of blood engenders on the luminal vessel wall a frictional force per unit area which is known as hemodynamic shear stress.

Hemodynamic forces regulate blood vessel structure and influence development of vascular pathology such as atherosclerosis, aneurysm, post stenotic dilation, and arteriovenous malformation [13, 14]. Vascular disease like Atherosclerosis remains one of the major cause of death in many countries. It is associated with genetic reason and numerous risk factors such as diabetes mellitus, hypertension, smoking, social stress, etc. Despite the nature of all its associated risk factors, atherosclerosis has an inclination to involve the outer edges of blood vessel bifurcation. In the vulnerable areas, blood flow is quite slow and it changes direction with cardiac cycle resulting in a weak hemodynamic shear stress. In contrast, vessel regions remain comparatively disease free that are exposed to steady blood flow and a higher shear stress.

Cerebral aneurysm is a blood-filled balloon-like bulge in the vessel wall of an artery located in the brain. Aneurysm can be thought to be developed due to weakness of the arterial wall. If it becomes large enough, it can rupture and spill blood into the adjacent tissue. Three types of cerebral aneurysms-Saccular, Fusiform and Giant are well known and separated by their geometry. The most common type of aneurysm is Saccular aneurysm which is of rounded shape and is attached to an artery by a neck or stem. It is also called as berry aneurysm as the shape is like berry. A Fusiform aneurysm which is a less common type of aneurysm, is of spindle-shape. Due to the widening of the vessel wall, it is formed. Another type of aneurysm, a Giant aneurysm, which is actually a berry aneurysm but it is large, occurs at the bifurcation of an artery.

There is no known prevention for aneurysm as the initiation of this vascular disease is not well understood. Various studies have been performed to provide the detailed hemodynamic information of the artery. Several studies have attempted to find out appropriate hemodynamic properties correlated with the aneurysm initiation. The hemodynamic properties such as the blood pressure, velocity of blood flow, and the wall shear stress are studied to be linked to the progression of the aneurysm [14, 15]. Wall shear stress is considered to be one of the hemodynamic reasons for the development of the cerebral aneurysm. The wall shear stress (wss) acts directly on the endothelium cell. Moreover, high wall shear stress or high spatial and temporal variation of wall shear stress (wss) may mechanically damage the inner wall of the artery [16, 17].

In this connection, the purpose of the work presented here is to study the CFD-based flow analysis and the effect of selected hemodynamic parameters on different cerebrovascular phantoms.

2 Preprocessing of 3-D Reconstruction for Flow Analysis

Development of appropriate 3-D mathematical model is one of the crucial tasks to be performed for thorough hemodynamic analysis of carotid vasculature. To do the hemodynamic analysis through the finite element modeling, we need synthetic 3-D structures, resembling human carotid arteries. It is informative to note that the 3-D

reconstruction involves two phases. During the first phase, we have designed a cerebrovascular phantom [4, 5] and then get a 3-D surface reconstruction for hemodynamic analysis. The second phase of the 3-D reconstruction process involves the conversion of the 3-D surface to a 3-D solid mesh. In case of complex structures, we have to process it through different preprocessing methods that discard spurious edges and vertices, etc. For generation of the initial surface mesh and for the rendering purpose, MeshLab_64bit_v1.3.4BETA is used [18]. To convert the surface mesh to solid mesh, we have used Rhinoceros 5.0 [19]. Figure 1a shows the initial surface mesh of a sample ACA phantom.

Figure 1b shows the surface mesh of the same structure after edge and vertex reduction. The quality of the mesh has been improved through some methods in

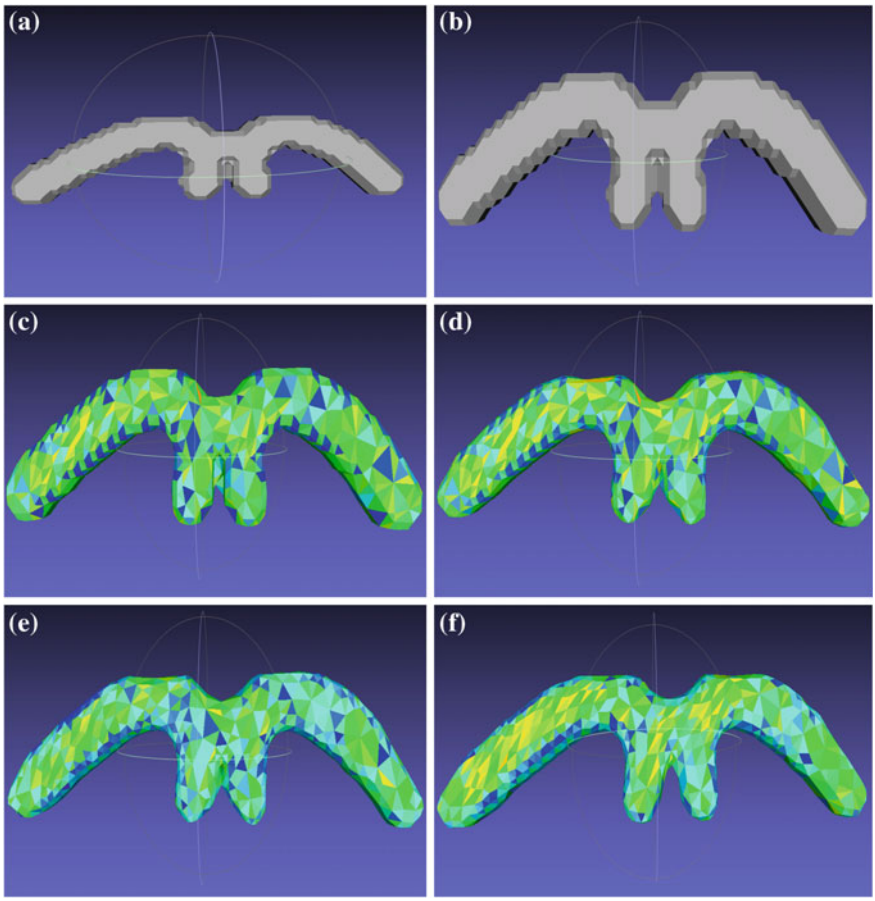


Fig. 1 Rendering of 3-D cerebrovascular phantom for ACA (Anterior Cerebral Artery) **a** initial mesh **b** after edge reduction **c** after MLS projection **d** after Laplacian smoothening **e** after Taubin smoothing **f** final mesh after resurfacing and mesh enhancement

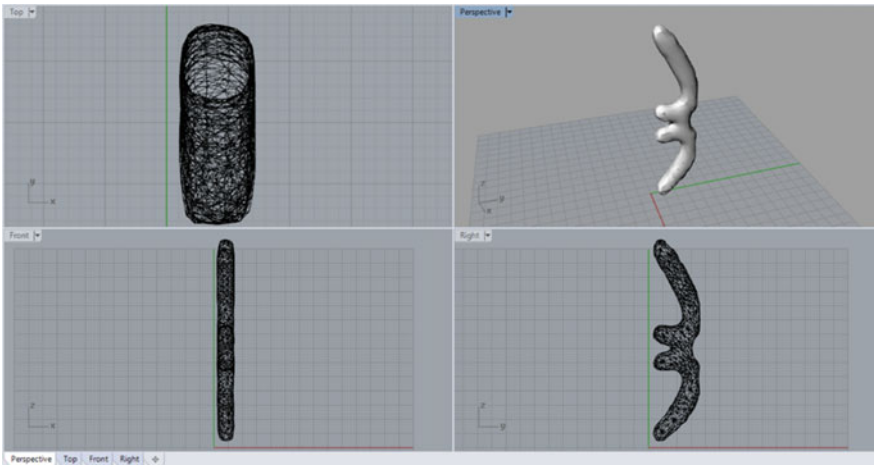


Fig. 2 Conversion from initial surface mesh to solid mesh in Rhinoceros 5.0

Meshlab 64bit. Figure 1c shows the moving least squares (MLS) projection of the phantom. In Fig. 1d and e, more preprocessing techniques have been applied to make the structure smoother. The quality of the mesh is represented using the color code, where "blue" regions represent best mesh quality. It may be observed, that the mesh quality has improved through different preprocessing steps applied on the structures, as shown in Fig. 1c and e. After all the smoothening operations, resurfacing method is applied over the phantom (see Fig. 1f).

Figure 2 shows the solid structure and surface mesh construction steps of the sample phantom. From the initial surface mesh, the solid mesh is generated using Rhinoceros 5. This shows how a digital cerebrovascular phantom (after conversion into surface mesh) can be converted to a solid mesh structure.

3 CFD-Based Flow Analysis

ANSYS [20] is a well-known computational fluid dynamics software and is used in our study for hemodynamic analysis on the cerebrovascular phantoms. It is widely used as a standard tool for finite element analysis. The fluent module of ANSYS Workbench 16.0 is used here to analyze different hemodynamic parameters through flow analysis. This analysis may be used as a benchmark to compare the digital flow based model, to be developed as a future direction of the current work.

The solid structure of the 3-D phantom is loaded in the design modeler of ANSYS is shown in Fig. 3a. Then the mesh is generated on the structure where inlet and outlets are specified in "Red" regions (see Fig. 3b). Now this mesh structure is ready for subsequent flow analysis. Another major component of the

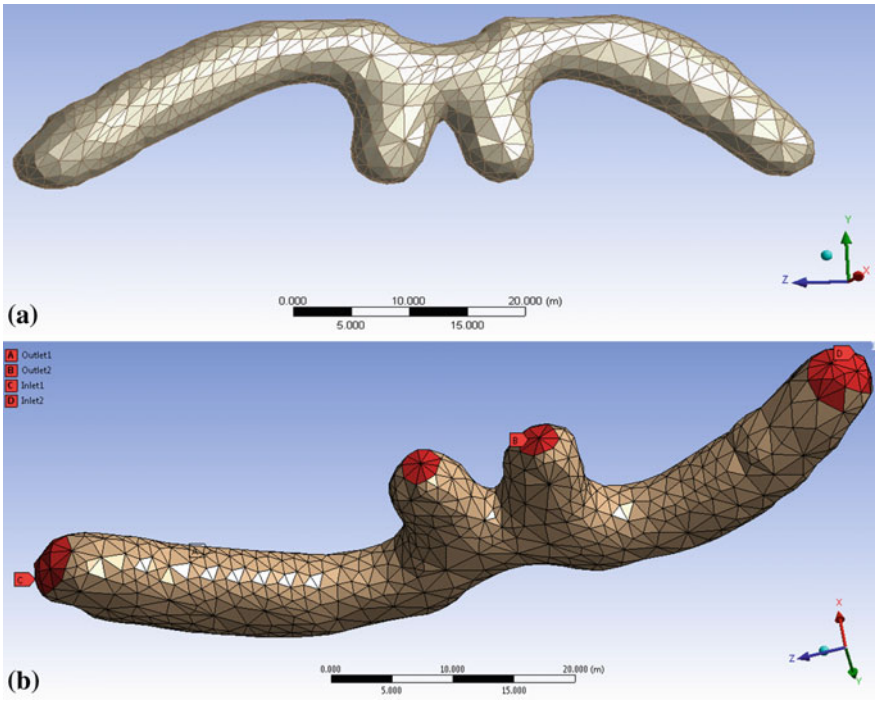


Fig. 3 Meshing of solid model. **a** Solid design model in ansys **b** meshing in ansys

cerebrovascular structure is the *circle of Willis*. This is shown in Fig. 4 as a part of the complete cerebrovascular phantom. Here, we show only the final result after converting the surface mesh to solid, without showing all the intermediate steps.

4 Results and Discussion

Velocity and wall shear stress are the major hemodynamic parameter which are analyzed in the present work. Our result using ANSYS Fluent software is shown in a contour filled way. For the numerical simulation of hemodynamic parameter in finite element method, we have used ANSYS16.0. Figure 5 shows the pulsatile flow through cerebrovascular phantom. In Fig. 5, we have pointed the inlets and outlets for this 3-D reconstructed digital phantom. In Fig. 6, the contour is filled with the velocity magnitude. In Fig. 7, we have shown the wall shear stress distribution for this phantom. The graphical representation of velocity magnitude is shown in Fig. 8.

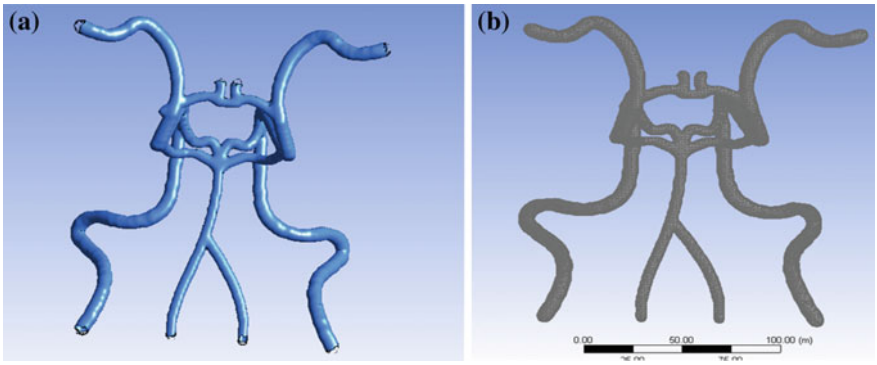


Fig. 4 **a** Phantom of major arteries of human cerebrovasculature along with *circle of Willis* **b** mesh

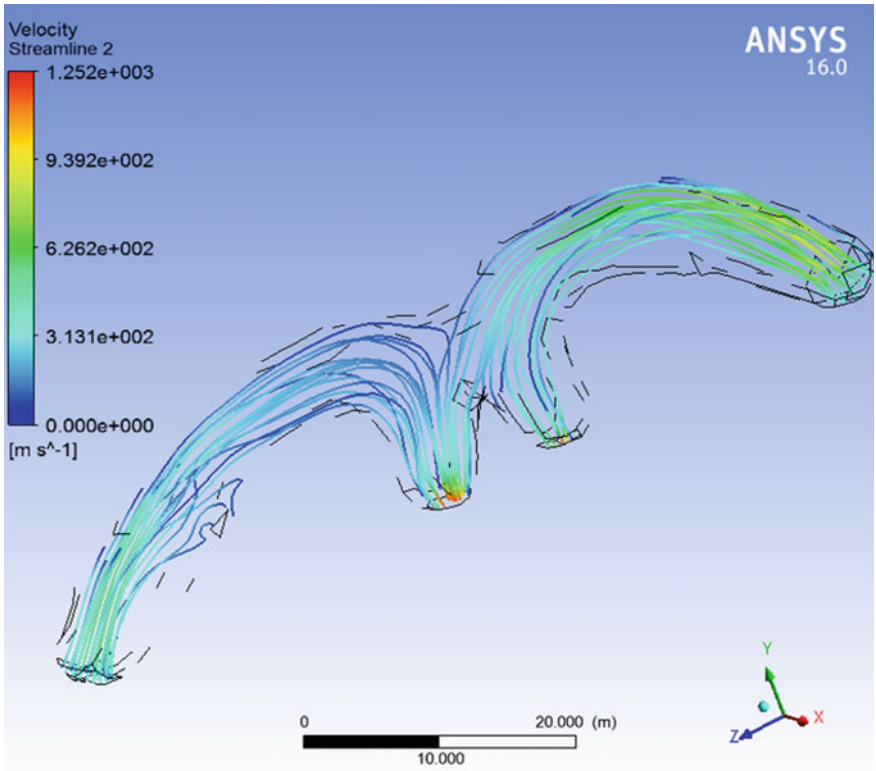


Fig. 5 Flow through cerebrovascular phantom (ACA)

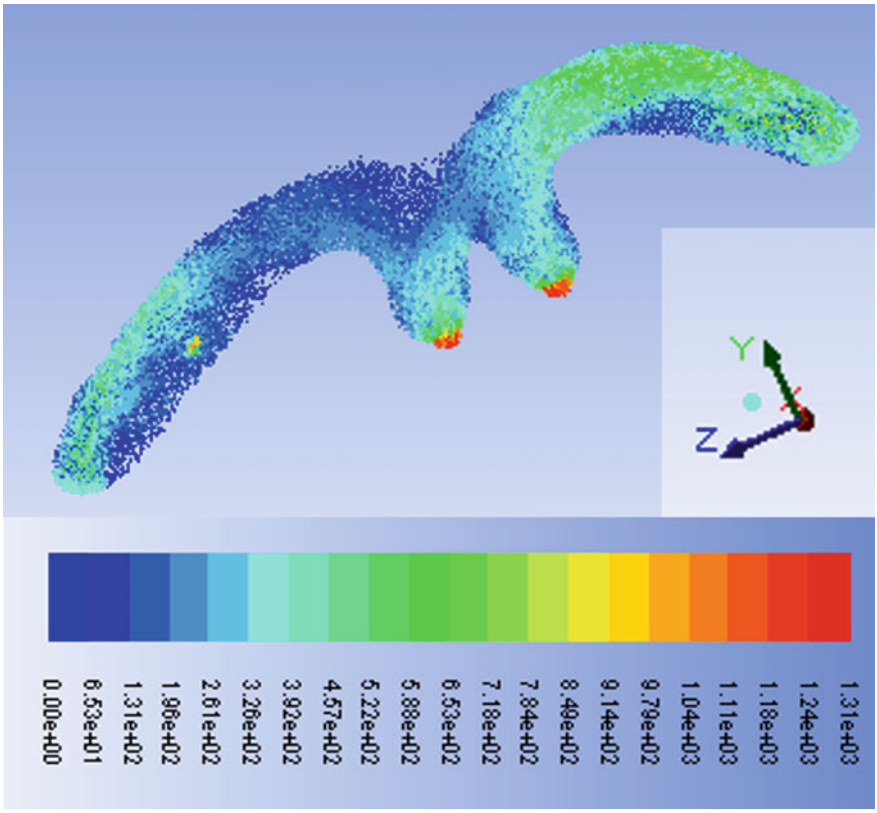


Fig. 6 Contour of velocity magnitude for ACA

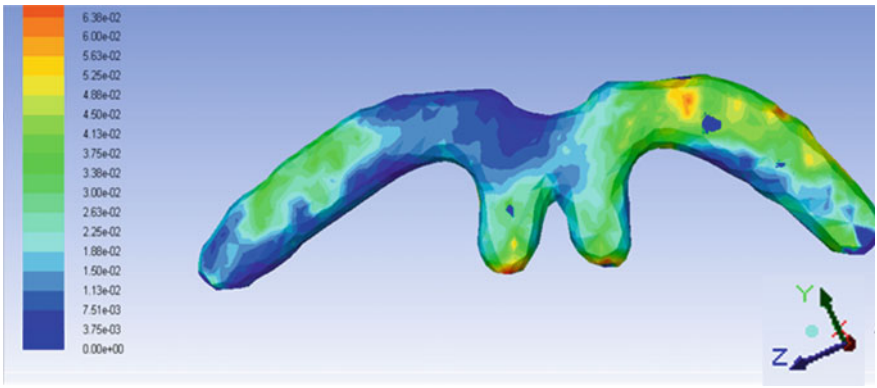


Fig. 7 Contour of wall shear stress (ACA)

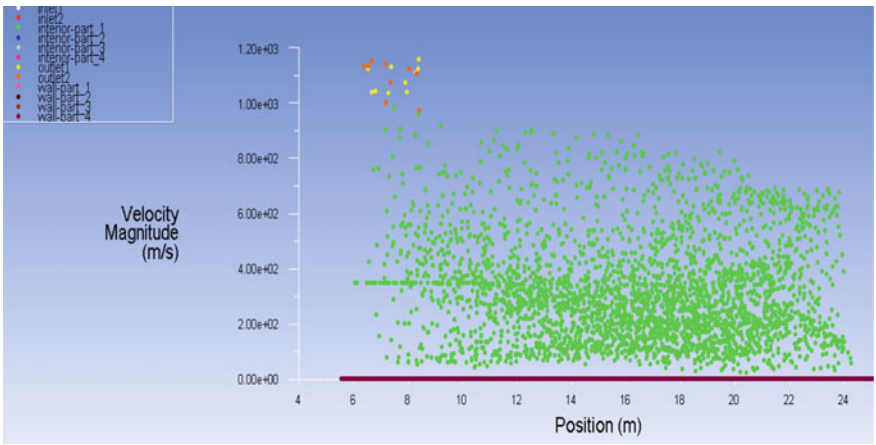


Fig. 8 Velocity magnitude (ACA)

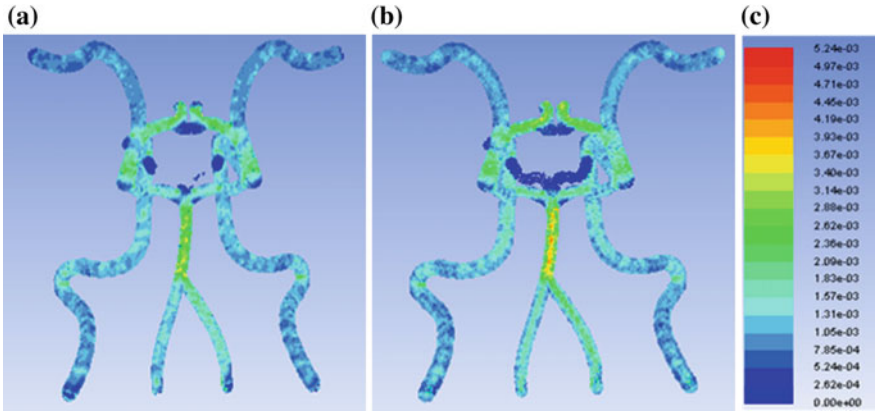


Fig. 9 **a** Contours of wall shear stress **b** velocity vectors (*Circle of Willis*) **c** transition of values with the transition from blue to red

Finally, in Fig. 9, we also show the result of flow analysis through the complete vasculature phantom which consists of major parts of carotid arterial structure including *circle of Willis*. Here, the lower four arterial regions are considered as inlets and the outlets are assumed in the upper sides of the arteries. The "Blue" regions show low velocity and shear stress in the vasculature, while the "Red" regions represent the highest velocity and shear stress, respectively.

5 Conclusion

In this present study, we have emphasized the utility of the reconstruction of cerebrovascular digital phantoms and the subsequent hemodynamic analysis through finite element method. At first, we have designed a ACA phantom then prepared it for flow analysis and then we have done the flow analysis through it using ANSYS Fluent CFD module. Secondly, we have designed and used another phantom which consists of major part of cerebrovasculature including *circle of Willis* for flow analysis. Analysis of fluid dynamic parameters reveals important structural properties. The color coded images represent both qualitative and quantitative maps of the digital phantoms. These results are useful for understanding of fluid flows through cerebral vasculature and serve as benchmarks for 3-D digital flow models through similar structures. Computational studies on digital flows are important future research directions for the current work. 2-D digital flows have already been used successfully for the study of structural/plastic changes in hippocampal dendritic spines [3]. The work presented may be included in the future study on the 3-D digital flow based hemodynamic analysis in both patients' CTA images as well as on complex mathematical phantoms.

Acknowledgements This work is supported by the DST PURSE-II, Government of India, project of CSE Department of Jadavpur University.

References

1. Gonzalez, C.F., Cho, Y.I., Ortega, H.V., Moret, J.: Intracranial aneurysms: Flow analysis of their origin and progression. Am. J. Neuroradiol. **13**(1), 181–188 (1992)
2. Vankan, W.J., Huyghe, J.M., Janssen, J.D., Huson, A., Hacking, W.J.G., Schreiner, W.: Finite element analysis of blood flow through biological tissue. Int. J. Eng. Sci. **35**(4), 375–385 (1997)
3. Basu, S., Plewczynski, D., Saha, S., Roszkowska, M., Magnowska, M., Baczynska, E., Wlodarczyk, J.: 2dSpAn: semiautomated 2-d segmentation, classification and analysis of hippocampal dendritic spine plasticity. Bioinformatics 172 (2016)
4. Banerjee, A., Dey, S., Parui, S., Nasipuri, M., Basu, S.: Design of 3-D phantoms for human carotid vasculature. In: Proceeding—2013 3rd International Conference on Advances in Computing and Communications ICACC 347–350 (2013)
5. Banerjee, A., Dey, S., Parui, S., Nasipuri, M., Basu, S.: Synthetic reconstruction of human carotid vasculature using a 2-D/3-D interface. In International Conference on Advances in Computing, Communications and Informatics (ICACCI) 60–65 (2013)
6. Saha, P.K., Wehrli, F.W., Gomberg, B.R.: Fuzzy distance transform: theory, algorithms, and applications. Comput. Vis. Image Underst. **86**(3), 171–190 (2002)
7. Saha, P.K., Udupa, J.K., Odhner, D.: Scale-based fuzzy Connected Image Segmentation: Theory, Algorithms, and Validation. Comput. Vis. Image Underst. **77**(2), 145–174 (2000)
8. Basu, S., Hoffman, E., Saha, P.K.: Multi-scale opening–a new morphological operator. In Image Analysis and Processing—ICIAP, Springer, pp. 417–427 (2015)
9. Saha, P.K., Strand, R., Borgefors, G.: Digital topology and geometry in medical imaging: a survey. IEEE Trans. Med. Imaging **34**(9), 1940–1964 (2015)

10. Basu, S., Raghavan, M.L., Hoffman, E.A., Saha, P.K.: Multi-scale opening of conjoined structures with shared intensities: methods and applications. In International Conference on Intelligent Computation and Bio-Medical Instrumentation (ICBMI), pp. 128–131 (2011)
11. Saha, P., Basu, S., Hoffman, E.: Multi-scale opening of conjoined fuzzy objects: theory and applications. IEEE Trans. Fuzzy Syst. (2015) (in press)
12. Guha, I., Das, N., Rakshit, P., Nasipuri, M., Saha, P.K.: Design of cerebrovascular phantoms using fuzzy distance transform based geodesic paths. In 4th International Conference on Advanced Computing, Networking, and Informatics (ICACNI) (2016) (in press)
13. Zarins, C.K., Giddens, D.P., Bharadvaj, B.K., Sottiurai, V.S., Mabon, R.F., Gladov, S., Glagov, S.: Carotid bifurcation atherosclerosis: quantative correlation of plaque localization with flow velocity profiles and wall shear stress. Circ. Res. **53**(4), 502–514 (1983)
14. Fillinger, M.F., Marra, S.P., Raghavan, M.L., Kennedy, F.E.: Prediction of rupture risk in abdominal aortic aneurysm during observation: Wall stress versus diameter. J. Vasc. Surg. **37** (4), 724–732 (2003)
15. Raghavan, M.L., Vorp, D.A.: Toward a biomechanical tool to evaluate rupture potential of abdominal aortic aneurysm: identification of a finite strain constitutive model and evaluation of its applicability. J. Biomech. 3**3**(4), 475–482 (2000)
16. Chien, A., Tateshima, S., Castro, M., Sayre, J., Cebral, J., Viñuela, F.: Patient-specific flow analysis of brain aneurysms at a single location: comparison of hemodynamic characteristics in small aneurysms. Med. Biol. Eng. Comput. **46**(11), 1113–1120 (2008)
17. Vorp, D.A., Raghavan, M.L., Webster, M.W.: Mechanical wall stress in abdominal aortic aneurysm: Influence of diameter and asymmetry. J. Vasc. Surg. **27**(4), 632–639 (1998)
18. "MeshLab." http://www.3d-coform.eu/index.php/tools/meshlab
19. "Rhino—Downloads." https://www.rhino3d.com/download. Accessed 12 Aug 2016
20. "ANSYS—Simulation Driven Product Development." http://www.ansys.com/

A Case Study for Ranking of Relevant Search Results

Rakesh Chandra Balabantaray and Santanu Ghosh

Abstract Now, we are flooded with data, yet we are starving for knowledge. The information we get is by using some platforms. Search engine acts as one of such platforms for getting information. So, when a user wants any information from the search engine, it should return some valid information. At first, the user gives his/her query to search engine. The query is matched with the documents present within the database of the search engine. This search engine uses different similarity functions to retrieve correct information from the database of the search engine. This similarity functions uses different mathematical ways to give a score. Now, the document from the database of the search engine which has the most significant score is retrieved. Like this, the results are retrieved and the user gets his/her information.

Keywords Search engine · Solr · BM25 · Default similarity

1 Introduction

Search engine is defined as set of programs that search relevant documents with respect to specified queries. These keywords are searched from database to return a list of documents where the keywords are found. Nowadays, the search engine term is used to describe systems like Google, Yahoo, etc., which searches relevant document from the World Wide Web. But actually they are web search engines.

R. C. Balabantaray (✉) · S. Ghosh
Department of Computer Science and Engineering, International Institute of Information Technology Bhubaneswar, Gothapatna, PO: Malipada, Bhubaneswar 751003, Odisha, India
e-mail: rakesh@iiit-bh.ac.in

S. Ghosh
e-mail: ghosh.santanu56@gmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_23

Solr enables us to create text based search engine which searches crawled websites, databases, and different files (like pdf, word, etc.). Now for getting crawled websites, we have to crawl using Nutch. It is a extensible open source web crawler software. It crawls the websites, while Solr indexes the crawled data available from Nutch. Now when any information is searched using Solr, the relevant information is provided from the indexed data.

Similarity in search engine is a known term used for searching relevant items from the collection of documents present within the search engine with respect to a user given query. The search engine actually calculates a score for the query and document. The document with most score is returned first. There are different types of similarity functions used by the search engine. Like Solr uses similarity like Default Similarity, BM25 similarity, etc. These similarity functions are obtained from different retrieval models. Such as, default similarity in Solr uses Vector Space Model [1]. The BM25 Model provides us the BM25 similarity.

2 Related Work

A lot of work is done in text-based search engine. Solr has been used widely for doing this work. Vector Space Model and BM25 [2] have also been used for research purposes. We have thought of using BM25 and Vector Space Model. We have studied few relevant articles for making of the documents.

The first paper describes about Process of Full Text Search Engine [3, 4]. It consists of steps like Building a Text Database, Create Indexing, Searching, Filtering, and sorting Results. It then says how the Full Text Search Engine works. It defines the five modules of Lucene. The first one comes as org.apache.lucene.analysis Analyzer. Its segments the document and removes the stop words. The second one that follows is org.apache.lucene.document. It does Document Management. The third one comes as org.apache.lucene.index. It establishes index. It also inserts and deletes records. Next it tells about org.apache.lucene.search. It tells about searching results of the queries given by Lucene. The fifth one is about org.apache.lucene.queryParser. It parses the user queries and returns the result to the user. It then explains about their system implementation. When they go to experimental results, they say that their result on basis of time consuming Lucene is better than String Retrieval system.

The next paper tells about BM25 and BM25F using Lucene Java Framework [5]. It then tells about mathematical formulas of BM25 and BM25F. The implementation part follows. It says at the Scorer level the functionalities are implemented, because weight and query will create necessary parameters for Scorers. The Scorers instances will be generated when the search is invoked. The execution of query is divided into two parts, a Boolean Filtering and documents ranking. Next it tells about how to use BM25 and BM25F. It says that BM25 or BM25F parameters

should be set before query is executed. This is done to get the average length and other parameters are got as it is got using default similarity.

In this paper, they say about construction of Inverted Index and building up a search engine using Apache Lucene [6]. They say about the data abstraction used by an index engine is an inverted index. They say about existing retrieval systems, like they are static. But they have presented data structures that efficiently realize the inverted index and permit continuous online application of updates without disturbing retrieval performance. Later they proceed on saying about their system architecture, where they have used web crawler to get data from web. Next, they have Lucene API for performing indexing and retrieving results from the queries given by the user. They have connected different processors using LANs, where each processor has separate disk. But the external user would see it as a single machine working in background.

This paper is about the vector space model [7]. The vector space model actually calculates the similarity score between text documents and queries in a term-document space. Next it proceeds to the experiment they conducted. First, in the experiment, it tells about how Vector Space Model works. It tells the documents are broken into word's frequency table. These tables are called vectors. Then vocabulary is built from the words present in all the documents of the system. A document and user query is recognized as a vector against the vocabulary. From there, the similarity is calculated. At last the documents are ranked for relevance. They have given an example of the score is calculated using cosine similarity. Next they proceeded by saying the variation available in vector space model. They mentioned about stop words, Synonyms, Proximity, Hyperlink, Position of word, etc [8].

This paper talks about the measures for search engine evaluation that are proposed and tested [1, 9]. The first measure they stated is Recall. Recall says the total amount of relevant items retrieved to the total number of relevant items in the collection. Recall is hard for calculation as it needs information about all the relevant items in collection. Next, they say about precision. Precision tells about the relevant items retrieved with respect to total items retrieved. Next, they state about relevance judgments with respect to humans. The humans decide whether the things retrieved were relevant or not. Next, measure they talk about the stability measure. In this stability measure, there is a computer program which gives relevancy judgment other than human beings. They used three websites for doing this experiment. They used single words search, phrase search, two word searched connected by Boolean AND, phrase search with word search joined by Boolean AND. Then the result was provided based upon the correlation among the search engine ranking and the human ranking. Next, they provided the percentage of top ranked pages retrieved.

This paper starts by saying the necessity of search engine today [10]. There are lots of places people go on searching in different online forums like social media, health forums, etc. The previous work on search engine is said. It started by saying about Google. Google has many crawlers. It crawls the data from the link it visits. Each page has a unique document id. The pages are compressed and stored. The indexer parses the page and turns them into word occurrences called hits, sorts

them, and then distributes them and at last they create a forward index. There is an anchor file which stores the related information about links stored. Google analyze the links available and ranks them. So when the time comes, it returns the right pages. This is Google's Page Ranking. Google's database is built in such a way that it can index and search large amount of pages taking a very less time. Then, it says about Google's big files and repositories. Next, they say about other search engines available one of them they mentioned is semantic search engine [11], which tells about sellers and service providers and their products that can be hierarchically organized. They talk about few search engine built over years. It says about Meta Search Engine, where results are combined by multiple search engines and present if required in a sorted order. They say about JSoup API, which is an HTML parser. It has developed a crawler for crawling the web and downloading web pages. Further they proceed saying about Indexing and Searching done by Apache Lucene. It is Java based but can also be used by programming languages like Perl, Python, and Net. Lucene uses efficient searching algorithms. Next, they have given few lines of code snippet for JSoup API to visit all links in a page, code snippet for Lucene for Indexing, code snippet for Searching using Lucene, code snippet for calculating scores by Lucene.

The paper of this author compares the relevance feedback on statistical language model also on binary independence probabilistic model [12–14]. The algorithm relevance feedback for traditional vector space model and binary independence probabilistic model is well developed. But the Binary Independent Probabilistic Model provides partial ranking for retrieved documents [12]. But an extension from binary independence model provides full ranking for documents [15]. These algorithms are known as best match algorithm. Then it compares the relevance feedback algorithm, which says that LM-algorithms shows significant improvement for retrieval performance than the BM25 algorithm for values of $b = 0.75$ and $k = 1.2$.

3 Experimental Setup

UNIX, JDK 1.7 and Apache Ant are used to perform the experiment.

3.1 Solr

Solr helps us to build a search engine which can search websites, files, and databases. For working with it, we have to download Solr and install it. Now to check whether it is working or not we can check it as.

We run Solr on web browser by initiating commands in command terminal as "SOLR_RUNTIME_HOME/example > java -jar start.jar". Then open web browser and we run the URL "http://localhost:8983/solr/#/".

3.2 Nutch

Nutch is extensible open source web crawler software. For working with it, we have to download nutch and install it. After installing, we have to check whether the installation is done rightly or not by the command "bin/nutch" that is to be run from ${NUTCH_RUNTIME_HOME}.

4 Optimization of BM25 Algorithm

BM25 stands for Best Match 25. In the below, the formula is shown.

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D) \cdot (k_1+1)}{f(q_i,D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)}, \tag{1}$$

f(q,D) represents q's term frequency in document D. |D| tells the size of the document D in words. queryNorm(q) is normalizing factor that make scores among queries comparable.

$$\text{q} = 1/\text{sqrt}(\text{sum Of Squared Weights}) \tag{2}$$

avgdl says about average document length in text collection from where documents are brought. k1 and b represents the free parameters. n represents total number of documents in collection. n(q) represents number of documents having q.

The free parameters can be used to get an optimized score and relevant results by changing those parameters. The experiment performed for optimizing BM25 algorithm is shown below in steps.

STEP-1

For performing this, we have crawled using Nutch and indexed using Solr. I have got 59361 documents for performing this experiment (Table 1).

STEP-2

After this, we have used twenty different queries. The queries are:

These queries are used and a maximum score is achieved for BM25 Similarity (b = 0 and k1 => 2).

STEP-3

Table 1 Queries

Jagannath puri	Obama	Jnu	India	Consultancy
IIIT Bhubaneswar	Lofoya	Espn	Narendra Modi	Congress
Delhi	Kashmir	Number series	Cat exam	Himalayas
Maths	Java	Rabindranath Tagore	Percentage	Temple

We have then checked the URLs whether the results shown were relevant or not. For this, we have used three measures as follows (Table 2).

After this, we have calculated the precision value based on P@3 (Precision value of top 3 URLs), P@5 (Precision value of top 5 URLs) and P@10 (Precision value of top 10 URLs) for different values of b and k1, and also for default similarity.

Then, we have calculated the Minimum Average Precision Value for Default Similarity and BM25 similarity (b = 0 & k1 = 0, b = 0 & k1 = 2, b = 1 & k1 = 0, b = 1 & k1 = 2, b = 2 & k1 = 0, b = 2 & k1 = 2).

5 Results and Discussions

5.1 Default Similarity

The Default Similarity uses Vector Space model where Cosine Similarity is used. The precision values are calculated with respect to the results returned on using Default Similarity (Table 3).

Table 2 Ratings

Relevant	Partially relevant	Nonrelevant
1	0.5	0

Table 3 Showing Precision Value for Default Similarity

Query	P@3	P@5	P@10
Jagannath puri	0.67	0.7	0.80
Obama	0.5	0.5	0.6
Jnu	0.5	0.4	0.45
India	0.5	0.5	0.4
Consultancy	0.83	0.8	0.9
IIIT Bhubaneswar	1	1	1
Lofoya	1	1	1
Espn	1	1	1
Narendra Modi	0.67	0.4	0.5
Congress	0.17	0.3	0.3
Delhi	0.5	0.4	0.35
Kashmir	0.17	0.2	0.15
Number series	0.3	0.4	0.4
Cat exam	0	0.17	0.05
Himalayas	0.5	0.5	0.65
Maths	0.5	0.7	0.55
Java	0.67	0.5	0.55
Rabindranath Tagore	0.5	0.6	0.45
Percentage	0.67	0.6	0.6
Temple	0.83	0.8	0.85

Table 4 Showing Precision value for b = 0 and k1 = 2

Query	P@3	P@5	P@10
Jagannath puri	1	1	1
Obama	1	1	1
Jnu	0.83	0.8	0.85
India	1	1	1
Consultancy	1	1	1
IIIT Bhubaneswar	1	1	1
Lofoya	1	1	1
Espn	1	1	0.95
Narendra Modi	0.83	0.8	0.8
Congress	1	1	1
Delhi	1	1	1
Kashmir	1	1	1
Number series	0.83	0.8	0.5
Cat exam	0.17	0.3	0.15
Himalayas	1	1	0.8
Maths	0.67	0.7	0.75
Java	1	1	1
Rabindranath Tagore	0.83	0.7	0.65
Percentage	1	1	0.9
Temple	1	1	1

5.2 BM25 Similarity (B = 0, K1 = 2)

The precision values are calculated for BM25 Similarity where b = 0 and k1 = 2 (Table 4).

We have got the best results for these values of b = 0 and k1 => 2.

5.3 BM25 Similarity (B = 2, K1 = 2)

The precision values are calculated for BM25Similarity where b = 2 and k1 = 2 (Table 5).

6 Comparison

The comparative analysis for different values of b and k1 of BM25 and Default Similarity, are presented in the figure given below: See Fig. 1.

Table 5 Showing Precision value for b = 2 and k1 = 2

Query	P@3	P@5	P@10
Jagannath Puri	0.67	0.6	0.75
Obama	0.5	0.5	0.6
Jnu	0.5	0.4	0.45
India	0.5	0.5	0.4
Consultancy	0.83	0.8	0.9
IIIT Bhubaneswar	1	1	1
Lofoya	1	1	1
Espn	1	1	1
Narendra Modi	0.67	0.4	0.4
Congress	0.17	0.3	0.3
Delhi	0.5	0.4	0.35
Kashmir	0.17	0.2	0.15
Number series	0.3	0.4	0.4
Cat exam	0	0.17	0.05
Himalayas	0.5	0.5	0.65
Maths	0.5	0.7	0.55
Java	0.67	0.5	0.55
Rabindranath Tagore	0.5	0.5	0.55
Percentage	0.67	0.6	0.6
Temple	0.83	0.8	0.85

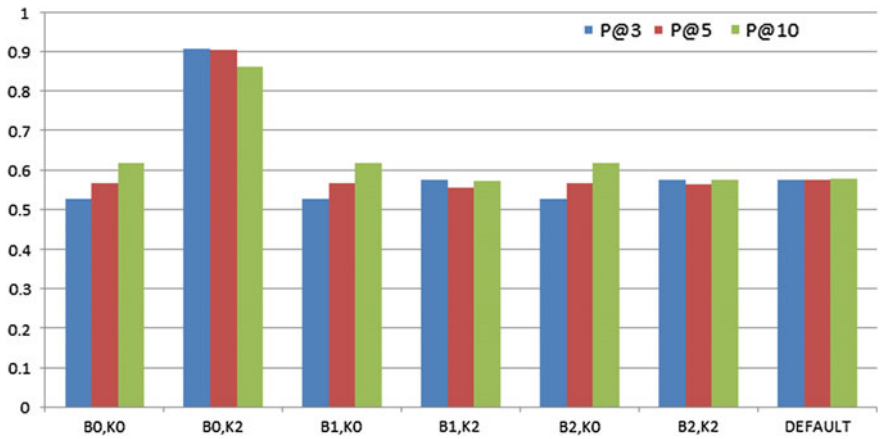


Fig. 1 Comparison between the different values of b and k1 of BM25 and Default Similarity, for P@3, P@5, P@10 is shown in the graphical form

7 Conclusion

Ranking and evaluation of search engine is a wide area of research. Ranking of relevant search results was shown in the comparison part. We have worked on that and got maximum relevancy and score for BM25 (b = 0, k1 => 2).

References

1. Raghavan, V.V., Wong, S.M.: A critical analysis of vector space model for information retrieval. J. Am. Soc. information Sci. **37**(5), 279 (1986)
2. Bollegala, D., Matsuo, Y., Ishizuka, M.: Measuring semantic similarity between words using web search engines. www **7**, 757–766 (2007)
3. Gao, R., Li, D., Li, W., Dong, Y.: Application of full text search engine based on lucene. Adv. Internet Things **2**(04), 106 (2012)
4. McCandless, M., Hatcher, E. Gospodnetic, O.: Lucene in action: covers apache Lucene 3.0. Manning Publications Co (2010)
5. Prez-Iglesias, J., Prez-Agera, J.R., Fresno, V., Feinstein, Y.Z.: Integrating the probabilistic models BM25/BM25F into Lucene (2009). http://arxiv.org/abs/0911.5046
6. Veer, G., Rathod, P., Sinare, P., Singh, R.B.: Building inverted index and search engine using apache Lucene. Int. J. Adv. Res. Comput Commun. Eng. **4**, 426–428 (2015)
7. Singh, V.K., Singh, V.K.: Vector space model: an information retrieval system. Int. J. Adv. Eng. Res. Studies/IV/II/Jan.–March, 141–143 (2015)
8. Hchsttter, N., Lewandowski, D.: What users seeStructures in search engine results pages. Inf. Sci. **179**(12), 1796–1812 (2009)
9. Vaughan, L.: New measurements for search engine evaluation proposed and tested. Inf. Process. Manage. **40**(4), 677–691 (2004)
10. Balipa, M., Balasubramani, R.: Search engine using apache lucene. Int. J. Comput. Appl. **127**(9), 27–30 (2015)
11. Dhyani, D., Bhowmick, S.S., Ng, W.K.: Deriving and verifying statistical distribution of a hyperlink-based Web page quality metric. In: Database and Expert Systems Applications, pp. 19–28. Springer, Berlin Heidelberg (2002)
12. Mukhopadhyay, D., Biswas, P., Kim, Y.C.: A syntactic classification based web page ranking algorithm (2011). http://arxiv.org/abs/1102.0694
13. Lv, Y., Zhai, C.: Adaptive term frequency normalization for BM25. In Proceedings of the 20th ACM international conference on Information and knowledge management, pp. 1985–1988. ACM (2011)
14. Hiemstra, D., Robertson, S.E.: Relevance feedback for best match term weighting algorithms in information retrieval, pp. 37–42 (2001)
15. Haveliwala, T.H.: Topic-sensitive pagerank: a context-sensitive ranking algorithm for web search. IEEE Trans. Knowl. Data Eng. **15**(4), 784–796 (2003)

Constrained Team Formation Using Risk Estimation Based on Reputation and Knowledge

Gaganmeet Kaur Awal and K. K. Bharadwaj

Abstract Given a generalized task that requires a different number of experts for various skills, the team formation problem (TFP) in real-world social networks aims to identify a set of experts that have the requisite skills and can collaborate effectively to accomplish the desired task. This paper considers TFP in realistic settings where the team composition must satisfy certain constraints. Sometimes for a task, only certain suitable experts having high reputation in the team of experts is sufficient to achieve the task. Moreover, not all experts having high reputation/ high expertise are always needed or are available for the task. To evaluate this, we propose a genetic algorithm-based model and introduce risk estimation strategies to determine the suitability of team for a particular task. The experimental results establish that our proposed model is useful for TFP in practical scenarios and discovers more coherent and collectively intelligent teams having low inherent risks.

Keywords Team formation · Genetic algorithm · Risk estimation
Reputation · Social networks

1 Introduction

The existence of web-based social networks (SNs) presents a collaborative platform to perform various complex tasks. Nowadays, the tasks have become so complicated that they require a multitude of diverse skills which is difficult for a single person to master and exhibit. Thus, this requires a collective expertise of more than one expert, i.e., forming a team of experts. This necessitates the development of automated algorithms for facilitating the process of team formation for such tasks.

G. K. Awal (✉) · K. K. Bharadwaj
School of Computer and Systems Sciences, Jawaharlal Nehru University,
New Delhi, India
e-mail: awal.gaganmeet@gmail.com

K. K. Bharadwaj
e-mail: kbharadwaj@gmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_24

The team formation problem (TFP) in the context of SNs is introduced by Lappas et al. [7] and after that, a significant amount of research has been done in this regard considering the communication costs of the team [6–8]. One such attempt has been made to measure the collective intelligence (CI) of the team by modeling the knowledge competence and collaboration competence of experts [1]. Several models for team formation have been proposed in the literature [1, 3, 6–8, 11] but no effort has been made to take risk estimation explicitly due to constraints posed by the tasks into account.

Reliability and trustworthiness are indeed deciding pillars for the success of any collaborative tasks. To reflect this aspect, our work considers the reputation of experts which is a well-known and widely used notion across domains like e-commerce, information retrieval, web-based services, etc. [2]. Only the Wisdom of the Crowds is not always sufficient for discovering collectively intelligent teams; sometimes, many scenarios may also require uncovering the Wisdom in the Crowds for achieving better performance [10].

There are numerous tasks, for example, the space mission critical tasks often require highly reputed experts; business-critical tasks require a diverse set of experts where entrepreneurs may also be needed to bring out innovative ideas; crowdsourcing tasks are usually not too complex and where experts are recruited individually and for which single reputed expert per subtasks are sufficient to achieve the goals of the task. Therefore, there are situations where an optimal allocation of experts to tasks with constraints needs to be done.

To this intent, reputation, as well as knowledge, serves as an essential filtering criterion to model various realistic constraints for TFP in real-world settings. To model the formation of teams under various constraints and to correctly quantify the collective abilities of the team members is a challenge in itself. To deal with this aspect of TFP, a genetic algorithm (GA) based model and risk estimation strategies are proposed that quantifies the implied risk.

The goal of this article is to determine the inherent risk associated with the process of forming teams and to show how various constraints can be modeled and characterized using risk analysis. Our contribution is threefold. First, we compute the reputation of experts from past time-stamped ratings on various projects. Second, GA is employed to discover teams at different sparsity levels while considering the various realistic constraints. Finally, to assess the inherent risk associated with discovered teams, we introduce estimation of risk at both global and local levels.

The rest of the paper is organized as follows: Section 2 summarizes the related work about TFP in SNs, CI, and reputation and also describes the problem formulation. Section 3 explains our proposed model for constrained TFP and introduces the risk estimation strategies. Section 4 provides our experimental results and analysis. Finally, the last section presents our conclusions and some research directions.

2 Preliminaries

2.1 Related Work

The TFP in SNs gained prominence in recent years due to the availability of social platform that allows experts to communicate and work together on projects. The work of Lappas et al. [7] is the first attempt to solve TFP in the SNs with the aim of minimizing the communication cost function (i.e., diameter and weight of minimum spanning tree) of the team which is improved in [6]. The TFP is extended for generalized tasks (i.e., the kind of task that requires a specific number of experts for each desired skill) by Li and Shan [8] by introducing density-based measures.

Wi et al. [11] modeled the capability of a team by considering knowledge of experts in addition to collaboration between the experts. To leverage the collective power of the members of a team and to model the synergistic aggregation of individuals' potential, Awal and Bharadwaj [1] attempt to assess the CI of the team by defining a quantitative measure, namely, collective intelligence index (CII). CII is based on enhanced expertise score and the collaboration score derived through trust between the experts. In this work, we use CII as the objective for forming the teams which are subjected to various constraints.

CI holds multidisciplinary relevance and is defined as the combined intelligence that emerges from the association and cooperation of individuals [10]. The coherent emergence at the global level that dynamically originate from interactions at the local level is studied in [9] by utilizing trust and reputation for communities of practice. The reputation can be computed according to various methods, e.g., a simple summation of ratings (eBay), an average of ratings (Amazon.com, Epinions), a weighted average of ratings [5], or fuzzy beta reputation models [2], etc.

2.2 Problem Formulation

Given an expertise SNs, the problem of organizing a team for the generalized task is to discover a group of experts such that: (i) for each requisite skill, the desired number of experts are allocated; (ii) all the requisite skills for the task are covered by the experts; (iii) the collective ability of the team is maximized; and (iv) all the constraints of team formation are satisfied.

3 Proposed Realistic Team Formation Model

In this work, our proposed model addresses the TFP for generalized tasks considering various realistic requirements. Usually, the goal of forming a team is to find a set of reputed and highly knowledgeable experts who can collaborate effectively

to accomplish a given task. But sometimes, the presence of few or more experts satisfying the above criteria is sufficient for the task. Our formulation allows modeling of many real-life constraints based on reputation and knowledge which are described below:

Type 1: Based on the reputation of experts.

- **Case A**: All experts should have reputation greater than the required threshold
- **Case B**: At least one expert per skill should have reputation greater than the required threshold
- **Case C**: Exactly one expert per skill should have reputation greater than the required threshold.

Type 2: Based on the expertise of experts.

- **Case D**: At least one expert per skill should have reputation greater than the required threshold as well as his/her expertise should lie in top k%
- **Case E**: Forming the best possible configuration of team such that for every skill in the generalized task, experts with maximum expertise in that particular skill are selected, and also all such experts should have reputation greater than the required threshold
- **Case F**: Forming the best possible configuration of team such that for every skill in the generalized task, experts with maximum expertise in that particular skill are selected.

3.1 Reputation Computation

Reputation plays a key role in effective interactions and determines the influence and importance of an expert over other team members. Trust is the short term behavior of an individual and is the local direct one-to-one information about another individual whereas reputation is the accumulative assessment of the long-term behavior and is a global phenomenon. Therefore, both trust and reputation are central to cooperative behavior between experts and also represents the phenomenon of CI by portraying its global-local property.

It represents commonly held belief about an agent's trustworthiness based on past interactions with other members of the network, and therefore, is a collective measure. It can be computed by using ratings-based approach adapted from [5] from the time-stamped interaction log.

The trust-based reputation value $T_{b->a}$ from expert E_b to expert E_a:

$$T_{b->a} = \frac{\sum_{\forall R_k \in R(E_b, E_a)} \phi_{I_k} \times \nu_k}{\sum_{\forall R_k \in R(E_b, E_a)} \phi_{I_k}}, \tag{1}$$

where $R(\mathrm{E}_b, \mathrm{E}_a)$ is the set of ratings given by E_b to E_a, v_k represents the rating's value for the rating R_k on the interaction I_k and ϕ_{I_k} gives the corresponding weight. The decay function for weight to the rating value is adapted from [1] so as to give more importance to recent ratings.

The proposed model then aggregates these values from all the experts in the network to derive a final reputation value Rep_a for an expert E_a.

$$\mathrm{Rep}_a = aggregate(\mathrm{T}_{b->a}, \mathrm{T}_{c->a}, \ldots, \mathrm{T}_{n->a}). \tag{2}$$

3.2 Knowledge Competence and Collaboration Competence

Knowledge competence (KC) and collaboration competence (CC) are important features that ascertain the efficiency and effectiveness of teams. KC is defined as "a measure of the leverage that team members have by virtue of being connected to and having interacted with a team of experts, in addition to their respective skills [1, 11]."

We determine the KC based on the expertise of experts and their strength of relationships and compute CC as the trust between them. The KC of an expert which is represented as "*enhanced expertise score* (EES)" is given by [1]:

$$\xi_x^{'} = \xi_x + \forall y \text{ s.t. } Adj(x,y) = 1 \left[\left\{ \max_{1 \le d \le \gamma; j = 1,2,\ldots,N} \mathrm{h}_d^j \times \left(\min_{\mathrm{E}_i, \mathrm{E}_{i+1}} sc_{i,i+1}^j ; \forall \mathrm{E}_i, \mathrm{E}_{i+1} \in Path_{xy}^j \right) \times \xi_y^j \right\} \right.$$
$$\left. \text{where } d = \left| Path_{xy}^j \right|, 1 \le d \le \gamma, \gamma = 3, 1 \le j \le N \right], \tag{3}$$

where $\xi_x^{'}$ and ξ_x are the EES and the personal expertise score of an expert $\mathrm{E_x}$, $Adj(x,y)$ corresponds to the adjacency matrix value for expert $\mathrm{E_x}$ and $\mathrm{E_y}$, γ is the maximum propagation distance, d is the distance for an expert from the source expert, h_d is the decay function over the propagation distance, sc_{ij} is the strength of connectivity among the experts $\mathrm{E_i}$ and $\mathrm{E_j}$ and $Path_{xy}^1, Path_{xy}^2, \ldots, Path_{xy}^N$ are N paths between experts $\mathrm{E_x}$ and $\mathrm{E_y}$.

The notion of trust quantifies the CC and is modeled as the "Interaction trust" between the experts and is given by Eq. 1. The max-min trust propagation mechanism is deployed to alleviate sparsity.

3.3 Modeling of Team Formation Problem Using GA

GAs are guided by the phenomenon of natural selection and mimic the Darwinian evolutionary principle of survival of the fittest. Because TFP is NP-hard [7], we propose GA-based model to find suitable optimal teams that also satisfy the constraints posed by the tasks. We have used a linear vector for chromosome

representation where an individual gene represents an expert. The number of partitions in the vector is equal to the number of skills required for the task. We use the two-point crossover and substitution mutation as genetic operators to produce new offspring.

The chromosomes generated must satisfy the constraints of the task. Thus, we have designed reparation strategy to ensure the feasibility of chromosomes at each stage of the GA process, i.e., repairing the chromosome such that the required constraints of the problem are met. For each of the cases representing various constraints, different reparation strategies are utilized. An elitist strategy is adopted that ascertains that the optimal solution generated so far in the evolutionary process is invariably preserved in the population. GA terminates when the maximum number of generations is reached.

Fitness Function. The fitness function measures the quality of chromosome solutions. It directs the evolutionary algorithm by permitting better chromosomes to breed and evolve and hence ameliorate the quality of solutions over succeeding generations. The KC and CC of the team are computed as the total sum of individual team members' enhanced expertise scores and summation of all pair-wise trust values between team members respectively. For forming teams, KC and CC are two essential components that represent team's CII, which is defined as:

$$CII = w_1 \times \xi_C + w_2 \times T_C, \tag{4}$$

where C is the candidate team, ξ_C, T_C represents the expertise score and the trust score of the team respectively, and w_1, w_2 are the weights for trust score and expertise score impact factors. Here, CII serves as the fitness function.

3.4 Risk Estimation

The purpose of forming teams for different applications has several risk requirements as well as diverse preferences for various constraints in the risk-CII balance. Therefore, the formulation of the risk strategies should be done in such a way that it can ease the formation process. The estimation of risk of a team is performed using the following proposed measures:

a. **RI:** ***Risk factor of the team***: It is a global measure and is given as:

$$\mathrm{RI} = 1 - \frac{\eta_T}{|T|}, \tag{5}$$

where η_T is the number of experts in the team that satisfies the given constraints and $|T|$ is the size of the team.

b. **RII:** ***Risk factor of the team based on skills***: It is a local measure and is given as:

$$\mathrm{RII} = 1 - \left(\frac{\frac{\eta_{s_1}}{|s_1|} + \frac{\eta_{s_2}}{|s_2|} + \frac{\eta_{s_3}}{|s_3|}}{3} \right), \tag{6}$$

where η_{s_1}, η_{s_2}, η_{s_3} are the number of experts that satisfy the given constraints in the skills s_1, s_2, s_3 of the team respectively, and $|s_1|$, $|s_2|$, $|s_3|$ are the total number of experts in the skills s_1, s_2, s_3 of the team, respectively.

c. **RIII:** ***Risk index based on skills***: It is also a global measure and is given as:

$$\mathrm{RIII} = 1 - \left(\frac{\delta_{s_1} + \delta_{s_2} + \delta_{s_3}}{3} \right), \tag{7}$$

where,

$$\delta_{s_i} = \begin{cases} 1; & \text{if at least one expert in skill } i \text{ satisfy the given constraints} \\ 0; & \text{otherwise} \end{cases}.$$

Therefore, the value of RIII lies in the set {0, 0.33, 0.67, 1}. If the value is zero, then the task is achievable and otherwise not.

3.5 Steps of the Proposed Model

Step 1 Compute the reputation for all experts in the network from the ratings given in the time-stamped interaction log

Step 2 For all the experts, derive their enhanced expertise scores and also compute the interaction trust values among them

Step 3 Discover suitable teams using GA-based constrained optimization

Step 4 Evaluate the risk estimates for teams obtained at various sparsity levels

4 Experiments and Results

We performed experiments on the synthetic dataset to demonstrate the effectiveness of our proposed model of forming teams in constrained settings.

4.1 Design of Experiments

We performed experiments on the expertise network comprising of 30 experts. Every expert is affiliated with only single skill out of the total three skills. A time-stamped interaction log of these experts is considered that spans across 50 tasks with varying

skill sets. The time window of the interactions among the experts lies between 10 and 500 days before the current date. G1, G2, G3, G4, and G5 are the five datasets taken with different sparsity levels, i.e., 90%, 85%, 80%, 75%, and 70% respectively. The parameters w_1, w_2, γ, and threshold are set to 0.5, 0.5, 3, and 0.4, respectively. The GA parameters settings are crossover probability (0.8), mutation probability (0.1), population size (20), and the number of generations (100).

GA is employed for the cases A, B, C, and D as mentioned in Sect. 3 and these schemes are referred to as $GA_{(A)}$, $GA_{(B)}$, $GA_{(C)}$, and $GA_{(D)}$. For the cases E and F, the experts from the network are selected that have the highest expertise and satisfy the condition mentioned in their respective constraints and are referred to as $Top_{(E)}$ and $Top_{(F)}$. The effectiveness of the proposed model is compared with the exhaustive search method (ESM). ESM is evaluated for all the four cases which are denoted as $ESM_{(A)}$, $ESM_{(B)}$, $ESM_{(C)}$, and $ESM_{(D)}$. To test the performance of our model, we utilize the proposed risk measures mentioned in Sect. 3.4.

4.2 Results

The risk associated with the discovered teams for different given constraints cannot be completely removed, but it can be analyzed and controlled. The results of risk assessment are represented using three measures RI, RII, and RIII. First, we compare the proposed GA-based model for cases $GA_{(A)}$, $GA_{(B)}$, $GA_{(C)}$, $GA_{(D)}$ with $ESM_{(A)}$, $ESM_{(B)}$, $ESM_{(C)}$, and $ESM_{(D)}$ respectively across various sparsity levels and the results are presented in Table 1. A comparison for the case $Top_{(E)}$ with $Top_{(F)}$ is also tabulated in Table 1.

A low value for risk indicates higher reliability on the team members and corresponds to their better performance. It can be observed that our GA-based proposed model considerably outperforms the ESM for all the cases across different sparsity levels. It is because the teams obtained through our model tend to satisfy the constraints and are suitable for practical scenarios. By using the proposed model, the risk associated with constrained TFP can be assessed effectively and efficiently. It renders substantial gains over the ESM approach that does not take the inherent risk into account for large real-world SNs. Moreover, using reputation as a screening criterion enhances the reliability and trustworthiness of team members for the task thus, justifying the need for using reputation as a collective measure at the global level.

Another observation from the Table 1 is that the RIII values for ESM method are not always zero for various sparsity levels across different cases. It signifies that in such cases, the task is unachievable as the subtask (/s) does not satisfy the given constraints, and the selected team members have high risk associated with them.

The performance in terms of CII values is shown in Fig. 1. As it can be observed quite intuitively, that the proposed constrained optimization models were able to mirror closely the optimal solution exploration done exhaustively, albeit with lower risk values. ESM is used as a baseline reference which we attempt to mirror, given

Table 1 Risk estimation for constrained team formation across various sparsity levels

Constraints	Sparsity levels	Methods					
Case A		$GA_{(A)}$			$ESM_{(A)}$		
		RI	RII	RIII	RI	RII	RIII
	G1	0	0	0	0.30	0.24	0
	G2	0	0	0	0.30	0.40	0.33
	G3	0	0	0	0.60	0.59	0
	G4	0	0	0	0.10	0.07	0
	G5	0	0	0	0.40	0.51	0.33
Case B		$GA_{(B)}$			$ESM_{(B)}$		
		RI	RII	RIII	RI	RII	RIII
	G1	0.20	0.18	0	0.30	0.24	0
	G2	0.10	0.17	0	0.30	0.40	0.33
	G3	0.50	0.52	0	0.60	0.59	0
	G4	0.10	0.07	0	0.10	0.07	0
	G5	0.40	0.37	0	0.40	0.51	0.33
Case C		$GA_{(C)}$			$ESM_{(C)}$		
		RI	RII	RIII	RI	RII	RIII
	G1	0	0	0	0.40	0.41	1
	G2	0	0	0	0.70	0.66	1
	G3	0	0	0	0.10	0.07	0.33
	G4	0	0	0	0.60	0.59	1
	G5	0	0	0	0.50	0.48	1
Case D		$GA_{(D)}$			$ESM_{(D)}$		
		RI	RII	RIII	RI	RII	RIII
	G1	0.70	0.60	0.33	0.80	0.67	0.67
	G2	0.60	0.59	0	0.70	0.76	0.33
	G3	0.70	0.66	0	0.70	0.66	0
	G4	0.60	0.49	0	0.60	0.49	0
	G5	0.60	0.54	0	0.90	0.89	0.67
Case E and Case F		$Top_{(E)}$			$Top_{(F)}$		
		RI	RII	RIII	RI	RII	RIII
	G1	0	0	0	0.50	0.57	0.33
	G2	0	0	0	0.40	0.41	0
	G3	0	0	0	0.60	0.59	0
	G4	0	0	0	0.30	0.29	0
	G5	0	0	0	0.50	0.58	0.33

that, for large real-world SNs, it is difficult to apply ESM. Moreover, $Top_{(E)}$ and $Top_{(F)}$ cases have relatively less CII values, which asserts that the performance of teams formed by taking only knowledgeable experts is low. Thus, it shows the importance of using both knowledge and trust as a measure for composing better and diverse teams that can collaborate effectively and efficiently and can accomplish the task.

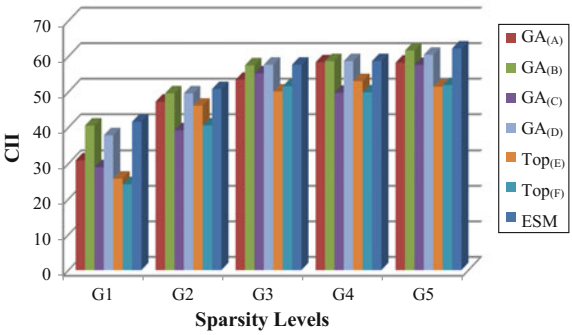


Fig. 1 Impact of various constraints on CII across sparsity levels

5 Conclusions

In this paper, we proposed the team formation problem (TFP) with real-world constraints in expertise social networks. We presented a model that helps in evaluating teams' performance subjected to various constraints in practical scenarios that are based on reputation and knowledge. Reputation serves as the global indicator to screen experts initially and after that, trust and knowledge are utilized for assessing the collective ability of the team. The experimental results have clearly indicated that our proposed model generates better teams at various sparsity levels with low-risk estimates than the teams obtained through exhaustive search method (ESM), thus, establishing the effectiveness of our proposed model. Moreover, teams formed by considering only knowledge criteria, i.e., by taking highly skilled experts (highest enhanced expertise scores) have comparatively fewer fitness values across sparsity levels than that of teams obtained through an evolutionary approach. It clearly demonstrates that diversity in the team leads to better overall performance; thus, proving the fact that knowledge, trust, and reputation are crucial factors for discovering teams with high collective intelligence (CI).

As future work, we plan to consider how experts' personality features [4] can be exploited to enhance the performance of teams. Another important direction would be to incorporate the concept of discovering team leader for the tasks [6]. A further extension of our approach could aim at incorporating availability [3] and workload of experts as additional constraints.

Acknowledgements This work is, in part, financially supported by Department of Science and Technology (DST), Government of India through the Inspire program.

References

1. Awal, G.K., Bharadwaj, K.K.: Team formation in social networks based on collective intelligence—an evolutionary approach. Appl. Intell. **41**(2), 627–648 (2014)
2. Bharadwaj, K.K., Al-Shamri, M.Y.H.: Fuzzy computational models for trust and reputation systems. Electron. Commer. Res. Appl. **8**(1), 37–47 (2009)
3. Dorn, C., Skopik, F., Schall, D., Dustdar, S.: Interaction mining and skill-dependent recommendations for multi-objective team composition. Data Knowl. Eng. **70**(10), 866–891 (2011)
4. Fan, Z., Feng, B., Jiang, Z., Fu, N.: A method for member selection of R&D teams using the individual and collaborative information. Expert Syst. Appl. **36**(4), 8313–8323 (2009)
5. Huynh, T.D., Jennings, N.R., Shadbolt, N.R.: FIRE: an integrated trust and reputation model for open multi-agent systems. In: 16th European Conference on Artificial Intelligence, pp. 18–22. IOS Press, Valencia, Spain (2004)
6. Kargar, M., Aijun, A.: Discovering top-k teams of experts with/without a leader in social networks. In: 20th ACM International Conference on Information and Knowledge Management, pp. 985–994. ACM, New York (2011)
7. Lappas, T., Liu, K., Terzi, E.: Finding a team of experts in social networks. In: 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 467–476. ACM, New York (2009)
8. Li, C-T., Shan, M-K.: Team formation for generalized tasks in expertise social networks. In: 2nd International Conference on Social Computing, pp. 9–16. IEEE, Washington (2010)
9. Maries, I., Scarlat, E.: Enhancing the computational collective intelligence within communities of practice using trust and reputation models. Trans. Comput. Collective Intell. **3**, 74–95 (2011)
10. Schut, M.C: Scientific Handbook for Simulation of Collective Intelligence. Available under creative commons license version 2 (2007)
11. Wi, H., Oh, S., Mun, J., Jung, M.: A team formation model based on knowledge and collaboration. Expert Syst. Appl. **36**, 9121–9134 (2009)

Compare Different Similarity Measure Formula Based Imprecise Query on Neutrosophic Data

Soumitra De and Jaydev Mishra

Abstract In this paper, we have designed two different similarity measure formulas which are applicable to find the closeness of two different neutrosophic data. After that compare the two different set of similarity measure values which are calculated by two different similarity measure formulas for better outcome. Then we have performed imprecise query on these two sets of similarity measure values and check which set based query will give better result for a certain tolerance value. Each neutrosophic data is based on truth, indeterminacy, and false membership values.

Keywords Neutrosophic data · Similarity measure formula · Indeterminacy query · Tolerance value

1 Introduction

Smarandache [1, 2] has discussed the logic of neutrosophic set in 2001, for managing the problems involving inconsistent data. A neutrosophic set is considered as a next generalization of vague set and is based on three interval-based memberships instead of vague set based two interval-based memberships. The three intervals value is based on neutrosophic data and is more appropriate for solving any indeterminacy query. Recent work on neutrosophic data and its applications are moving ahead rapidly and we have customized the neutrosophic database boundary according to our observation. A very few work on neutrosophic set has been reported in literature [3–6]. Here, we have designed two different similarity measure

S. De (✉) · J. Mishra
Computer Science & Engineering Department, College of Engineering and Management, Kolaghat, Purba Medinipur 721171, West Bengal, India
e-mail: soumitra@cemk.ac.in

J. Mishra
e-mail: jsm03@cemk.ac.in

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_25

formulas which are used on neutrosophic data for finding closeness value between two neutrosophic data separately. Several authors have used vague set [7–10] to execute imprecise query but no such work has been reported literature using similarity measure formula on neutrosophic data. Here, our objective is to run imprecise query on two different sets of similarity measure value for getting better query based outcome. Here each set of similarity measure value is calculated by different similarity measure formula (S.M). Imprecise query is based on SQL command with some tolerance value for retrieving desired tuples from the table of a database.

2 Neutrosophic Set

A neutrosophic set N on the universe of discourse (U) is given by:

- truth membership $t_n \rightarrow [0, 1]$,
- false membership $f_n \rightarrow [0, 1]$ and
- indeterminacy membership function $I_n \rightarrow [0, 1]$ such that $t_n + f_n \leq 1$ and $t_n + f_n + i_n \leq 2$ and is represented as $N = \{< n, [t_n, i_n, f_n]>, n \in U\}$.

3 Different Similarity Measure Formulas on Neutrosophic Set

Here, we have designed two different similarity measure formulas, which are an expansion of similarity measure for vague data [11, 12] to calculate closeness value between two neutrosophic data. The newly introduced similarity measure formulas are given below.

Similarity measure (S.M.) between two neutrosophic values:

Let m and n be any two neutrosophic values such that $m = [t_m, i_m, f_m]$ and $n = [t_n, i_n, f_n]$, where $0 \leq t_m + f_m \leq 1$, $0 \leq t_n + f_n \leq 1$, $t_m + i_m + f_m \leq 2$, $t_n + i_n + f_n \leq 2$.

Type-1 Formula

Let SE (m, n) denote the similarity measure between m and n.

Then,

$$SE(m,n) = \sqrt{\left(1 - \frac{|(t_m - t_n) - (i_m - i_n) - (f_m - f_n)|}{3}\right)(1 - |(t_m - t_n) + (i_m - i_n) + (f_m - f_n)|)}$$

Type-2 Formula

Let SE (m, n) denote the similarity measure between m and n.

Then,

$$SE(m,n) = 1 - \frac{|(t_m - t_n) + (i_m - i_n) + (f_m - f_n)|}{6} - \frac{\max\{|(t_m - t_n)|, |(i_m - i_n)|, |(f_m - f_n)|\}}{3}$$

4 Generate Similarity Measure Values Using Neutrosophic Data

In this work, we have experimentally observed how two neutrosophic data can retrieve different similarity measure values using two different formulas. These calculated similarity measure values are further required to run an imprecise query with certain tolerance value to examine which formula-based similarity values are optimized as per the closeness of tolerance value is concerned. To explain this, we have considered the following employee relational database as given in Table 1 and uncertain query given below.

4.1 Solution

At first, we have designed the Exp attribute into neutrosophic representation of Exp on which two different similarity measure formulas are used to find the closeness of two neutrosophic data. Next to apply imprecise query on two different similarities measured based columns with some tolerance value. Neutrosophic representation of Exp is shown in Table 2 on which similarity measure formulas are applied.

Table 1 Employee relation

EName	Exp (yrs)	Esal (Rs)
Jones	11	15000
Tomas	8	10000
Kates	10	12000
Joshep	12	20000
Rock	19	28000
Smith	14	24000
Rokey	16	26000
Adams	25	40000
James	20	30000
Doglus	13	23000

Table 2 Neutrosophic representation of exp in employee relation

EName	Exp (yrs)	Esal (Rs)	Neutrosophic representation of Exp [t, i, f]
Jones	11	15000	<11, [0.97, 0.03, 0.02]>
Tomas	8	10000	<8, [0.73, 0.15, 0.25]>
Kates	10	12000	<10, [0.96, 0.04, 0.03]>
Joshep	12	20000	<12, [1, 0, 0]>
Rock	19	28000	<19, [0.68, 0.35, 0.31]>
Smith	14	24000	<14, [0.95, 0.03, 0.04]>
Rokey	16	26000	<16, [0.91, 0.12, 0.06]>
Adams	25	40000	<25, [0.43, 0.42, 0.52]>
James	20	30000	<20, [0.65, 0.53, 0.30]>
Doglus	13	23000	<13, [0.98, 0.04, 0.01]>

We calculate the closeness values of Exp using Type-1 formula of similarity measure.

For example, let us consider the two neutrosophic data m = <12, [1, 0, 0]> and n = <11, [0.97, 0.03, 0.02]>.

Here,

$$t_m = 1, i_m = 0, f_m = 0, t_n = 0.97, i_n = 0.03, f_n = 0.02$$

Then

$$S.M.(m,n) = \sqrt{\left(1 - \frac{|(1-0.97)-(0-0.03)-(0-0.02)|}{3}\right) \quad (1-|(1-0.97)+(0-0.03)+(0-0.02)|)}$$
$$= \sqrt{0.973 \times 0.98} = \sqrt{0.9535} = 0.98$$

Again, for m = <12, [1, 0, 0]> and n = <8, [0.73, 0.15, 0.25]>, $t_m = 1, i_m = 0, f_m = 0, t_n = 0.73, i_n = 0.15, f_n = 0.25$.

This gives

$$S.M.(m,n) = \sqrt{\left(1 - \frac{|(1-0.73)-(0-0.15)-(0-0.25)|}{3}\right) \quad (1-|(1-0.73)+(0-0.15)+(0-0.25)|)}$$
$$= \sqrt{0.776 \times 0.87} = \sqrt{0.6751} = 0.82$$

Again, for m = <12, [1, 0, 0]> and n = <10, [0.96, 0.04, 0.03]>, $t_m = 1, i_m = 0, f_m = 0, t_n = 0.96, i_n = 0.04, f_n = 0.03$.

This gives

$$S.M.(m,n) = \sqrt{\left(1 - \frac{|(1-0.96)-(0-0.04)-(0-0.03)|}{3}\right) \quad (1-|(1-0.96)+(0-0.04)+(0-0.03)|)}$$
$$= \sqrt{0.963 \times 0.97} = \sqrt{0.9341} = 0.97$$

and so on.

Next, we calculate the closeness values of Exp using Type-2 formula of similarity measure.

Again, for m = <12, [1, 0, 0]> and n = <19, [0.68, 0.35, 0.31]>, $t_m = 1, i_m = 0, f_m = 0, t_n = 0.68, i_n = 0.35, f_n = 0.31$.

This gives

$$S.M.(m,n) = 1 - \frac{|(1-0.68)+(0-0.35)+(0-0.31)|}{6} - \frac{\max\{|(1-0.68)|,|(0-0.35)|,|(0-0.31)|\}}{3}$$
$$= \frac{5.66}{6} - \frac{0.35}{3} = 0.943 - 0.116 = 0.83$$

Again, for m = <12, [1, 0, 0]> and n = <14, [0.95, 0.03, 0.04]>, $t_m = 1, i_m = 0, f_m = 0, t_n = .095, i_n = 0.03, f_n = 0.04$.

This gives

$$S.M.(m,n) = 1 - \frac{|(1-0.95)+(0-0.03)+(0-0.04)|}{6} - \frac{\max\{|(1-0.95)|,|(0-0.03)|,|(0-0.04)|\}}{3}$$
$$= \frac{5.98}{6} - \frac{0.05}{3} = 0.996 - 0.016 = 0.98$$

Again, for m = <12, [1, 0, 0]> and n = <16, [0.91, 0.12, 0.06]>, $t_m = 1, i_m = 0, f_m = 0, t_n = 0.91, i_n = 0.12, f_n = 0.06$.

This gives

$$S.M.(m,n) = 1 - \frac{|(1-0.91)+(0-0.12)+(0-0.06)|}{6} - \frac{\max\{|(1-0.91)|,|(0-0.12)|,|(0-0.06)|\}}{3}$$
$$= \frac{5.91}{6} - \frac{0.12}{3} = 0.985 - 0.04 = 0.96$$

and so on.

The calculated two different similarity measures for the neutrosophic data are shown in Table 3.

Now, we perform the imprecise query with tolerance value on Type-1 and Type-2 based data set to see which type of similarity measure values will retrieve less numbers of tuples from the Table 3.

Next, we process the imprecise query on similarity measure values of neutrosophic data (Tables 4 and 5):

SELECT * FROM EMPLOYEE WHERE $T_1 \geq 0.95$

SELECT * FROM EMPLOYEE WHERE $T_2 \geq 0.95$.

Table 3 Two different S.M calculation for exp

EName	Exp (yrs)	Esal (Rs)	Neutrosophic representation of Exp [t, i, f]	S.M of Exp using Type-1(T_1)	S.M of Exp using Type-2(T_2)
Jones	11	15000	<11, [0.97, 0.03, 0.02]>	0.98	0.99
Tomas	8	10000	<8, [0.73, 0.15, 0.25]>	0.82	0.88
Kates	10	12000	<10, [0.96, 0.04, 0.03]>	0.97	0.98
Joshep	12	20000	<12, [1, 0, 0]>	1	1
Rock	19	28000	<19, [0.68, 0.35, 0.31]>	0.67	0.83
Smith	14	24000	<14, [0.95, 0.03, 0.04]>	0.97	0.98
Rokey	16	26000	<16, [0.91, 0.12, 0.06]>	0.91	0.96
Adams	25	40000	<25, [0.43, 0.42, 0.52]>	0.55	0.75
James	20	30000	<20[0.65, 0.53, 0.30]>	0.56	0.74
Doglus	13	23000	<13[0.98, 0.04, 0.01]>	0.97	0.98

Table 4 Imprecise query on T_1 with tolerance value = 0.95

EName	Exp (yrs)	Esal (Rs)
Jones	11	15000
Kates	10	12000
Joshep	12	20000
Smith	14	24000
Doglus	13	23000

Table 5 Imprecise query on S_2 with tolerance value = 0.95

EName	Exp (yrs)	Esal (Rs)
Jones	11	15000
Kates	10	12000
Joshep	12	20000
Smith	14	24000
Rokey	16	26000
Doglus	13	23000

5 Conclusion

A comparison between two different similarity measure formulas based on neutrosophic data and calculating similarity measure values using these formulas has been expressed in our current work. Each formula is used to find the closeness value set from two neutrosophic data. Here, one neutrosophic value is always fixed and other is changed tuple to tuple. Finally, SQL command based imprecise query is

executed with certain tolerance value on the similarity measure value based specific single column of a table. Here, we performed two different queries on two different columns of similarity measure value set. Finally, we observed that Type-1 formula based imprecise query retrieves less numbers of tuples than Type-2 formula based imprecise query and Type-1 formula based outputs are more closure than Type-2 based output. So, we are determined that Type-1 formula is more suitable for calculating similarity measure values from the table of a relational database which is based on neutrosophic data.

References

1. Smarandache, F.: First International Conference on Neutrosophy, Neutrosophic Probability, Set, and Logic, vol. 1. University of New Mexico (2001)
2. Smarandache, F.: Definitions derived from neutrosophics, multiple-valued logic. Int. J. **8** 591–604 (2002)
3. Arora, M., Biswas, R.: Deployment of neutrosophic technology to retrieve answers for queries posed in natural language. In: 3rd International Conference on Computer Science and Information Technology, vol. 3, pp. 435–439 (2010)
4. Arora, M., Biswas, R., Pandy, S.U.: Neutrosophic relational database decomposition. Int. J. Adv. Comput. Sci. Appl. **2,** 121–125 (2011)
5. Broumi, S.: Generalized neutrosophic soft set. Int. J. Comput. Sci. Eng. Inf. Technol. **3,** 17–30 (2013)
6. Deli, I., Broumi, S.: Neutrosophic soft relations and some properties. Ann. Fuzzy Math. Inf. **9**, 169–182 (2015)
7. Zhao, F., Ma, M.Z., Yan, L.: A vague relational model and algebra. Fourth International Conference on Fuzzy Systems and Knowledge Discovery, vol. 1, pp. 81–85 (2007)
8. Zhao, F., Ma, M.Z.: Vague Query Based on Vague Relational Model, vol. 1, pp. 229–238. Springer, Berlin (2009)
9. Mishra, J., Ghosh, S.: A new functional dependency in a vague relational database model. Int. J. Comput. Appl. **39** 29–33 (2012)
10. Mishra, J., Ghosh, S.: Uncertain query processing using vague sets or fuzzy set: which one is better? Int. J. of Comput. Commun. Control **9** 730–740 (2014)
11. Lu, A., Ng,W.: Managing Merged Data by Vague Functional Dependencies. LNCS vol. 3288, pp. 259–272. Springer, Berlin (2004)
12. Pei, Z., Liu, J.: Research on similarity measures between vague sets. In: Forth International conference on Fuzzy Systems and Knowledge Discovery, vol. 3, pp. 648–652 (2007)

Path Executions of Java Bytecode Programs

Safeeullah Soomro, Zainab Alansari and Mohammad Riyaz Belgaum

Abstract Static analysis of programs is essential for better understanding towards software maintenance and re-engineering. Unfortunately, we still lack automatic tools to understand the back end of the programs (Bytecode). Developing these tools is very expensive and time-consuming task but it is today's need. Those tools may help to understand Java Bytecode. Some time source code is not available all the time but bytecode is easily available. Unfortunately, bytecode is not understandable by many of us so that we are providing a little effort in this regard. This article represents the program flow execution in Java Bytecode. We present static and dynamic path executions of programs in a bytecode using Control Flow Graph (CFG) and Data Dependence Graph (DDG). Bytecode analysis is an effort to develop a tool which can make visualization of Java programs in back end form.

Keywords Static analysis of programs · Program dependence graph
Control flow graph · Software testing and maintenance

1 Introduction

Program execution is the process of control flow information towards output. It shows the behavior of programs in a dynamic way. Mainly, there are two ways of analysis of programs they are static and dynamic. Static analysis provides all text of programs and more path execution which can make some heuristics about programs. Dynamic analysis provides exact execution which is a dynamic path of the

S. Soomro (✉) · M. R. Belgaum
College of Computer Studies, AMA International University, Salmabad, Bahrain
e-mail: s.soomro@amaiu.edu.bh

M. R. Belgaum
e-mail: bmdriyaz@amaiu.edu.bh

Z. Alansari
University of Malaya, Kuala Lumpur, Malaysia
e-mail: zeinab@amaiu.edu.bh

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_26

program during runtime. In our paper, we are dealing with both approaches towards path executions. We present the Control Flow Graph (CFG) and Data Dependence Graph (DDG) from Java Bytecode which are better for developing tools of software testing and maintenance in future. Currently, we are lacking those tools which can provide information of whole programs in visual forms and that can be better for understanding of programs.

Program analysis is the process of verification of data flow and control flow information from programs [1]. This is although very active research area and many tools [2, 3] have been developed for the programs but unfortunately, all tools work on front end (Source Code) of programs. Mainly all tools are used for the source code but there are few tools [4], which have been developed for Java Bytecode for understanding and representation of data and control flow analysis. This area of research still needs more focused time so that people can get benefit from the back end (Bytecode) analysis which may reduce the cost of program maintenance and reengineering. At [5, 6], static analysis of Java Bytecode and Dependence analysis is presented which are essential for the understanding of computer programs. It may prove helpful for many software engineering tasks like testing, debugging, reverse engineering, and maintenance. Our approach presents new approach to building back end (Bytecode) tools for the Java Bytecode programs and may be helpful for upcoming automatic tools towards software testing, debugging, maintenance, and reverse engineering. The authors provided specification based model for the abstract dependencies from Java programs which helps towards finding faults and localizing faults [7–9]. The visualization of data through software requirement [7] is presented. To the best of our knowledge, our technique may help further to investigate program analysis and help towards debugging [10] in a bytecode.

This article makes use of static and dynamic analysis of programs in terms of bytecode analysis. In our approach, we have presented executions of programs, showing the Control Flow Graph (CFG) and Data Dependency Graph (DDG). A Control Flow Graph (CFG) presents the execution paths in a program using graph notations. It shows exactly the traversing of all statements of the program during execution. In a Control Flow Graph (CFG), the program statements are converted into nodes and edges, nodes show the basic blocks of the statements and edges show the control from one statement to another statement. There are, in most presentations, two specially designated blocks: the entry block, through which control enters into the flow graph, and the exit block, through which all control flow leaves. In our work, we are extracting edges and nodes from bytecode of the Java programs which helps in static and dynamic analysis of programs from the back end (Bytecode). Furthermore, we have presented Data Dependence Graph (DDG) from the bytecode which presents the constraints on how a piece of code can be reorganized in term of dependency. Furthermore, data dependency shows the relationship of the variables in a program.

The rest of this paper is organized as follow. Section 2 contains information of Java Virtual machine and Bytecode Information. Section 3 contains program execution information. Section 4 contains Control Flow Graph (CFG) from bytecode. Section 5 extracted Data Dependence Graph from the bytecode. Section 6 contains Related Research. Finally, conclusion and future research are depicted in Sect. 6.

2 Java Virtual Machine (JVM) and Bytecode Understanding

This section gives an overview about Java Virtual Machine (JVM) and bytecode information from the Java source code. The Java Virtual Machine (JVM) is a load based virtual machine which can support Java programming language [11]. It is an independent platform for the input of class files. Each class is a binary file which contains information about fields and methods. Java Programs are compiled into bytecode called as machine language of JVM which also provides opcode and operands in bytecode information [12]. The JVM is responsible for loading all relevant class files upon execution of program. At runtime, the JVM fetches opcode and corresponding operands and executes the corresponding actions accordingly [5]. Java Bytecode is an object code of Java program [13]. In a bytecode, it shows the line number, opcode and operands with complete information of variables used with reference numbers. Also, it provides prefixes to all statements used in source code of Java program. For example, if we used declaration of any data type in the source code then bytecode provides prefix with values assigned. For example *int* $i = 3$ represented in a bytecode is *ICONST*3, *i* declares the *integer* type of the variable and *CONST* shows that variable declaration and assigned value is 3. It also provides line number with *ISTORE* information, which is used for integer storage in the memory stack. So always in a bytecode, it includes the prefix information, line number, memory storage and label number which may count as line number in our idea. For basic and complicated statements of source code, compiler provides information in a bytecode with reference number, opcode, and corresponding operands.

3 Path Executions

This section contains information of the program execution and representation in control flow graph and data dependence graph from the back end (bytecode). There are two kinds of the execution of programs named as static and dynamic.

- **Static Execution:** It provides whole text of the program for analysis. Static always provides all information of program having all control flow possibility according to the source code. In Java program, we extracted all possible path executions. In the program, we found four path executions according to true and false values for those conditions. We have shown the static path executions of our program (Fig. 1) as follows:

```
Line Number 3: x > y : FALSE
Line Number 7: y > 5 : FALSE
Execution Path 1: 0 1 2 3 5 6 7 9 10
Line Number 3: x > y : TRUE
Line Number 7: y > 5 : FALSE
```

```
Execution Path 2: 0 1 2 3 4 6 7 9 10
Line Number 3: x > y : FALSE
Line Number 7: y > 5 : TRUE
Execution Path 3: 0 1 2 3 5 6 7 8 10
Line Number 3: x > y : TRUE
Line Number 7: y > 5 : TRUE
Execution Path 4: 0 1 2 3 4 6 7 8 10
```

- **Dynamic Execution:** It provides the exact flow control of program according to source code of program execution. It depends on compiler to compute and execute program statements based on the input values and other control flow statements of the program. We have presented dynamic execution path of our program (Fig. 1) as under:-

```
Line Number 3: x > y : FALSE
Line Number 7: y > 5 : FALSE
Execution Path : 0 1 2 3 5 6 7 9 10
```

4 Control Flow Graph

A Control Flow Graph (CFG) is a graph which represents control through whole program. It contains nodes of program which represents statements while edges show the flow of control between statements.

In the Fig. 1, an example of Java program is written and we have shown the execution passing through all paths. In Table 1 we have shown each bytecode statement of the Java Program, source code, basic blocks, and nodes for the graph. However, we have extracted source code and provided bytecode in the Table 1 is for understanding of the execution of Java programs. Our approach is to derive control flow graph and dependence graph from its bytecode. We have extracted all bytecode statements from the source code and have made blocks and nodes of all statements. Once program has been compiled, we analyzed bytecode. In Table 1, there are four columns, one column shows bytecode information, second represents source code, third column shows the basic blocks, and the last column shows the nodes for the graph. We recognize nodes according to entry and exit point of control flow execution of program.

Always we start from the first node and leader of the control flow graph. The leaders of all blocks have to be recognized through control flow from its entry and exit point. First, it begins with simple statements and making basic blocks for all the bytecode instructions. The basic block is defined as a block consisting of sequence of instructions where entry and exit point are only in one direction. So simple statement and multiple statements are sequentially treated as one node in the control flow graph. For example, we have made one node for all the sequence statements in our example

Fig. 1 Java test program

```
    public class TestProgram {
        public static void main(String[] args) {
0:          int x = 3;
1:          int y = 4;
2:          int z = 0;
3:           if (x > y)
4:              z = x + 2;
             else
5:              z = y + 2;
6:          z = z + y;
7:           if (y > 5)
8:              z = z + 5;
             else
9:              z = z - 2;
10:         z = z + 3;
        }
    }
```

program. In the Table 1, $B1$, $B2$ and $B3$ are counted as one node because of sequence of statements and no edges.

For those statements which have more than one edge then each statement is counted as a node. All conditional, calling methods statements in a program are counted individually and are assigned node for every edge in between. Each conditional instruction of program is counted as also leader so that we have given representation as node. If there is *if*, *while* or other comparative statements, also they are counted as leaders. The branch values of these leader are true and false. Each instruction of method calls and return also counted as leaders. After counting all instructions of programs we have assigned leaders and fixing the nodes. Once all bytecode instructions are represented in the form of nodes, the control flow graph works through all possible path executions. Also, we never count node of any return statements which may not return any value so we have reduced that block from the control flow, as it does not show impact on the flow of program and as well as result of the program.

Algorithm 1 Algorithm of Extracting Blocks and Nodes from Bytecode

INPUT : $BytecodeofJavaProgram(P)$
OUTPUT: $Nodes(N)andBlocks(B)$
Let Assume Line Number is L
Read P
while find *ICONST*
Add B
if $preL = nextL$
Add N
DrawGraph(B, N)
Ensure: {∀S∃P}

Table 1 Bytecode of program with blocks and nodes

Byte code	Source Code	Blocks	Nodes
L0			
LINENUMBER 0 L0			
ICONST3	x = 3	B0	N1
ISTORE 1			
L1			
LINENUMBER 1 L1			
ICONST4	y = 4	B1	N1
ISTORE 2			
L2			
LINENUMBER 2 L2			
ICONST0	z = 0	B2	N1
ISTORE 3			
L3			
LINENUMBER 3 L3			
ILOAD 1			
ILOAD 2			
IFICMPLE L4	if (x > y)	B3	N2
L5 LINENUMBER 4 L5			
ILOAD 1			
ICONST2			
IADD			
ISTORE 3	z = x + 2	B4	N3
GOTO L6			
L4			
LINENUMBER 5 L4			
ILOAD 2			
ICONST2 IADD			
ISTORE 3	z = y + 2	B5	N4
L6			
LINENUMBER 6 L6			
ILOAD 3			
ILOAD 2			
IADD			
ISTORE 3	z = z + y	B6	N5
L7			
LINENUMBER 7 L7			

(continued)

Table 1 (continued)

Byte code	Source Code	Blocks	Nodes
ILOAD 2			
ICONST5			
IFICMPLE L8	if (y > 5)	B7	N6
L9			
LINENUMBER 8 L9			
IINC 35	z = z + 5	B8	N7
GOTO L10			
L8			
LINENUMBER 9 L8			
IINC 3 −2	z = z - 2	B9	N8
L10			
LINENUMBER 10 L10			
IINC 33	z = z + 3	B10	N9
L11			
LINENUMBER 11 L11			
RETURN	return from method	B11	N10

In Fig. 2, for the building CFG from the bytecode, we use nodes which are created from blocks. We are not considering *else*, *if*, *return*, {, } and those which cannot make impact on the output of the programs. We are extracting nodes from the basic blocks from Table 1 and did not even count some basic blocks which are making from nonstructuring statements like *else*, *if*, *return*, {, *and*}. Our Algorithm 1 is to recognize the line numbers from bytecode and nodes which are created from blocks. After that we follow path execution of program in bytecode statically and dynamically as we have discussed in Sect. 3. We have presented below static and dynamic execution of paths according to nodes representation in the Fig. 2 as follows:-

- **Path 1**: N1, N2, N4, N5, N6, N8 and N9
- **Path 2**: N1, N2, N3, N5, N6, N8 and N9
- **Path 3**: N1, N2, N4, N5, N6, N7 and N9
- **Path 4**: N1, N2, N3, N5, N6, N7 and N9

In the above information path, 1, 2, 3, and 4 are static paths. Also, path 1 is counted as dynamic path because of compiler execution according to values in the program. We have extracted all information from bytecode and built blocks and nodes from program lines. After that nodes and blocks are presented in CFG which is shown in Fig. 2.

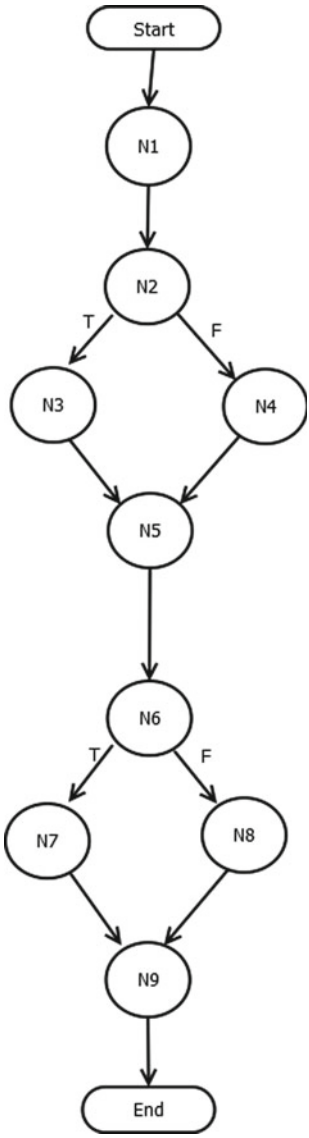


Fig. 2 Control flow graph of bytecode execution

5 Data Dependence Graph (DDG)

The Data Dependence Graph (DDG) is derived from assignment of variables which shows dependence relation in between. We extract information from bytecode and make blocks of each lines. We have given names to each block and counted dependencies of it. One block to another block we consider variable which may impact on another variable in a block. We have inspected each block in a Java Bytecode in

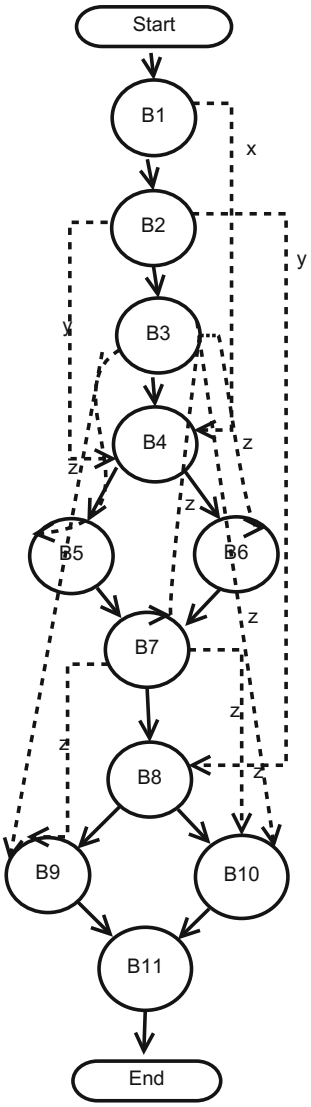


Fig. 3 Data dependence graph of bytecode execution

terms of values in variables. In Java code, $x = 3$ which represents $L0$ LINENUMBER ICONST 3 ISTORE 1 in bytecode. We calculated data dependencies according to given line number of each bytecode statements. Furthermore identify the names of variables, line numbers and uses of those variables in another block. We have identified the data flow of variables from one block to another block in a graph 3.

In Fig. 3, we have made blocks from all statements including simple, multiple and conditional. Blocks consist of all lines with variables and values. The simple lines show the control flow of the program and dotted lines show the dependency of

variables on different blocks. For example in the block number 3, variable z is used which can change in others blocks due to data dependency. We have shown in the Fig. 3 all variables dependencies according to blocks wise.

6 Related Research

In [14], a library that enables bytecode transformations by strategic rewriting has been presented using the language TOM. Mapping of bytecode programs to algebraic terms is done. Pattern matching and strategic programming primitives to the existing language is added to express bytecode transformations.

In [15], Fixpoint algorithms was used for analyzing the bytecode considering a number of optimizations in order to reduce the number of iterations. The term parametric is used as the algorithm is independent of abstract domain and it can be applied to different domains.

In [16], authors carried research work to discussed challenges faced by bytecode analyzers. With various example programming statements, the relation between low level and high level analyses using the concepts of strong and weak relative completeness have been formalized.

In [17], a framework for java program analysis called Soot was discussed. Various features of the framework were discussed which can be used for program analysis.

We have presented analysis technique and show how we can extract bytecode information from the code and represent it in control flow graph and data dependence graph. It is being good to understand the back end of the program. As we cannot find much material on understanding of bytecode our research may help to understand bytecode information.

7 Conclusion and Future Research

We have presented static and dynamic execution of programs from Java Bytecode instructions. It is essential for better understanding towards software maintenance and reengineering. Our article represents the program flow executions in Java Bytecode. We presented Control Flow Graph (CFG) and Data Dependence Graph (DDG) from bytecode information. We believe that our discussion and idea may help researchers to develop advance tools for understanding back end code of programs.

Future research has to develop tool of our presented idea and hope it may help for software debugging and testing community in future, which is really today's need in the world.

References

1. Sreedhar, V.C.: Efficient Program Analysis Using DJ Graphs. Doctoral Dissertation. McGill University, Canada (1995)
2. Java Checker. http://www.gradsoft.ua/products/javachecker_eng.html
3. Static Analysis Tools Exposition (SATE). http://samate.nist.gov/SATE.html
4. Dr. Garbage. http://www.drgarbage.com
5. Zhao, J.J.: Static analysis of bytecode. Wuhan Univ. J. Natural Sci. **6**(1–2), 383–390 (2001)
6. Zhao, J.J.: Dependence analysis of java bytecode. In: Proceeding COMPSAC 24th International Computer Software and Applications Conference, pp. 486-491. IEEE Computer Society Washington, DC, USA (2000)
7. Soomro, S., Abdul, H., Syed, H.A.M., Asadullah, S.: Ontology based requirement interdependency representation and visualization. In: Communication Technologies, Information Security and Sustainable Development Communications in Computer and Information Science vol. 414, 2014, pp 259–270, pp. 486–491. CCIS Springer Series (2013)
8. Soomro, S., Wotawa, F.: Detect and localize faults in alias-free programs using specification knowledge. In: LNAI Springer Series , IEA/AIE 2009, LNAI 5579, pp. 379–388 (2009)
9. Soomro, S.: Using abstract dependences to localize faults from procedural programs. In: Proceedings Artificial Intelligence and Applications, Innsbruck, Austria, pp. 180185 (2007)
10. Weiser, M.: Programmers use slices when debugging. Communications of the ACM **25**(7), 446452 (1982)
11. Arnold K., Gosling, J.: The Java Programming Language, Addision Wesley (1996)
12. Don, L., Roland, H.U., Nancy J.W.: Bytecode-based java program analysis In: Proceedings of the 37th Annual Southeast Regional Conference , ACM Southeast Conference, Mobile, AL, April 15–18 (1999)
13. Peter, H.: Java Bytecode (2001). https://www.ibm.com/developerworks/ibm/library/it-haggar_bytecode/
14. Ballad, E., Moreau, P.E., Rellies, A.: Bytecode rewriting in Tom. In: Second Workshop on Bytecode Semantics, Verification, Analysis and Transformation–Bytecode 07 Braga/Portugal (2007)
15. Maendez, M., Jeorge, N., Hermenegildo, M.V.: An Efficient, Parametric Fixpoint Algorithm for Analysis of Java Bytecode. Published in Electronic Notes in Theoretical Computer Science (2007). https://www.elsevier.nl/locate/entcs
16. Logozzo, F., Fahndrich, F.: On the relative completeness of bytecode analysis versus source code analysis. In: Published in 17th International Conference, CC 2008, Held as Part of the Joint European Conferences on Theory and Practice of Software, ETAPS 2008, Budapest, Hungary, March 29–April 6 (2008)
17. Lam, P., Bodden, E., Lhotak, O., Henden, L.: The Soot Framework for Java Program Analysis: a Retrospective. Published in Cetus Users and Compiler Infastructure Workshop ETUS (2011)

An Approach to Track Context Switches in Sentiment Analysis

Srishti Sharma and Shampa Chakraverty

Abstract Ever-increasing social media platforms provide colossal amounts of opinionated data. Sentiment analysis on social media is a valuable tool for the process of understanding these new means of online expression, detecting the relevant ones, and analyzing and exploiting them appropriately. Through this work, we introduce an innovative approach for separating text that conveys more than one theme. It focuses on efficiently segregating these different themes, sometimes known as context switches, and then accurately mining the different opinions that may be present in the text containing context switches. We utilize three categories of features namely positional, lexical semantic, and polarity features for theme-based text segmentation within a document. Themes of all the segments are obtained by using a simple noun phrase extractor and sentiment analysis on the different segments is performed to extract the opinions. We also propose an application that improves the efficiency of sentiment analysis and illustrates its working on two sample opinionated documents.

Keywords Opinion mining · Sentiment analysis · Social media · Context switches · WordNet · SentiWordNet

1 Introduction

Our beliefs and the choices we make, are often influenced by how people around us perceive and gauge them. Therefore, individuals as well as organizations seek out the opinions of others while making day-to-day and crucial decisions. In the

S. Sharma (✉) · S. Chakraverty
Department of Computer Engineering, Netaji Subhas Institute of Technology
New Delhi, New Delhi, India
e-mail: srish.060788@gmail.com

S. Chakraverty
e-mail: apmahs.nsit@gmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_27

information age, social networking platforms like Quora, Facebook, Twitter as well as product review websites help us gather the views of our friends and colleagues. Marketers, psychologists, and others view these websites as a repository for exploration and extraction of ideas. They provide the ability to monitor real-time feedback. This in turn has propelled a rapid surge in the demand for computational tools having the power to automatically extract and analyze relevant opinions from text. As a result, the research community is abuzz with new works on the extraction of sentiments from blogs, product review websites, and social media platforms.

Oftentimes, humans may differ while adjudging the sentiment expressed by a text. This demonstrates how complex sentiment analysis (SA) can get. The shorter the string of text to mine the opinion from, the tougher the job becomes. More often than not, the information contained in a product review, a blog or a microblog typically discusses one central idea or theme. As a result, most of the SA approaches focus on extracting data relevant to a particular theme and then mining the opinions from the collected dataset.

Humans have a natural flair for drawing upon past experiences and knowledge for presenting their ideas effectively. This results in a gradual flow of various sub-themes that converge to convey the main ideas. A core challenge is to automatically detect the switches between various contexts that occur within a text. In this work, we develop an innovative method to segregate various contexts effectively and then use Sentiment Analysis (SA) to mine the opinions from each of them.

2 Related Work

Several researchers have delved into the field of SA recently. In the beginning, the focus was on SA at the document level. SA at document level is aimed at categorizing any presented document into positive or negative polarities [1, 2]. With the growing prominence of micro-blogging, a more relevant problem is to carry out SA at the sentence level. A number of researchers have explored SA on Twitter in order to effectively classify tweets into positive, negative or neutral [3–5]. SA on sentences presents the problem of comparative sentences wherein two or more items are pitted against each other. The authors in [6] discuss the identification of comparative sentences in text. In [7], the authors present a technique for mining the preferred entities from comparative sentences. The works in [8–10] discuss opinion mining at the feature level. It entails extracting different features of an object and correctly identifying authors' sentiments corresponding to each of them.

Detection and segregation of sentiments corresponding to different entities from a piece of text continues to remain a challenge. A piece of text may contain more than one different idea corresponding to separate entities. Through this work we address the handling of these context switches, where the discussion abruptly changes from one theme to a completely different or somewhat related theme and

efficiently mining sentiments from them. Our work takes a different route from prior works as we tackle the problem of detecting context switches and utilize it to refine opinion mining.

3 Proposed Work

The proposed scheme for dealing with context switches in SA comprises of Theme Co-referent Text segmentation, Segment-Wise Theme Detection and Segment-Wise SA. We discuss each of them in detail.

3.1 Theme Co-referent Text Segmentation

In a text document containing context switches, there may be more than one theme being discussed. Hence, for efficiently dealing with context switches, the identification of number of different ideas present in text and their effective segregation from one another is crucial.

Co-referent segments are segregated by computing the similarity between pairs of sentences and subsequently, clustering similar ones. For clustering, we use Hierarchical Agglomerative Clustering (HAC), a bottom-up clustering approach. At the onset, HAC considers each item to be a cluster in itself and then ensues by successively merging or agglomerating cluster pairs. This continues till all the clusters are merged into one. As the number of clusters is not known beforehand, HAC is used on all the sentences in a text document and the clusters obtained from this hierarchy are used to identify the different segments.

The accuracy of pairwise co-reference clustering depends on the feature set chosen. In this work, we propose a feature set comprising of six features that fall under three different categories as discussed below:

Positional Features. Humans have a natural tendency to express one idea before switching over to another idea and then probably to some other idea and so on. Therefore it is only reasonable to consider that two sentences closer to each other in text have a higher probability of being theme co-referent that those farther apart. Two positional features are introduced to capture this effect:

Consecutive Sentences. We record whether the two sentences being analyzed are consecutive to one another.

Number of Intervening Sentences. If the two sentences are not consecutive, than record the number of intervening sentences between the two to analyze how far apart the two sentences under consideration are to each other.

Lexical and Semantic Features. Three lexical semantic features are introduced to group together co-referent sentences in text. For this set of features, stop words and punctuations are removed and stemming is performed. Stop words are removed

so that words like a, an, the, etc., that appear in almost all sentences and are not indicative of the semantics of the sentence are not taken into account. Stemming is performed to convert every word into its root form. Since no stemmer has 100% accuracy, two different stemmers are used namely, Porter Stemmer and Lancaster Stemmer [11].

Term Frequency-Inverse Document Frequency (TF-IDF) based Cosine Similarity. A vector is derived from each sentence in a document cosine similarity between every pair of sentences is recorded. Each position in the vector relates to a word in the vocabulary and it contains the TF-IDF of that word. TF-IDF is a frequency based indicator which signifies the importance of a word to a document in a corpus of documents [12]. Cosine similarity between two TF-IDF vectors for two texts indicates how alike the two sentences are in terms of their subject matter.

Word Overlap. Word overlap records whether the two sentences being analyzed have any overlapping words. The rationale behind recording word overlap is that theme co-referent sentences have a high probability of having certain content words which are likely to be repeated amongst sentences pertaining to the same theme.

WordNet Synonyms. Record whether the two sentences being analyzed have any words that may be synonyms of each other. For this purpose, the Lexical ontology WordNet is utilized [13]. It is relevant to check for synonyms because generally while discussing an idea, an author tends to use similar words.

Polarity Features. It is reasonable to assume that theme co-referent sentences have the same polarity, as an author will have one view on a theme either positive or negative and hence this is used as a feature for text segmentation.

Sentiment Polarity. We record whether two sentences under consideration have the same sentiment polarity. Before checking for polarity, Word Sense Disambiguation (WSD) is performed to check for the meaning of each word in a document and assign a sense number to every word as listed in WordNet. The classic Lesk algorithm [14] is used for WSD. Polarity of sentences is computed using Sentiwordnet [15]. The sentiment scores assigned by SentiWordNet to any word are in the range 0 to 1 where the sum of all three scores for every word is unity. The positive and negative scores of a word *w*, represented as *pos_score(w)* and *neg_score(w)* respectively, are extracted from SentiWordNet. If both these scores for a word are zero, the word is labeled as objective *obj* and its score is set to zero. If positive score is greater than negative score for the word, the word is positive and its label is set to *pos* and its positive score is taken as its score. Otherwise, the word is negative, labeled *neg* and its negative score is taken as the score. The polarity of a sentence is calculated by subtracting the polarities of words labeled *neg* from polarities of words labeled *pos*. The words labeled *obj* are ignored. If the final polarity of the sentence comes out to be positive, the sentence is labeled *pos*, otherwise it is labeled *neg*.

After extracting these features, the system applies HAC on all sentences in a given text document to generate clusters of sentences with cohesive themes.

3.2 Segment-Wise Theme Detection

Once a piece of text is divided into different theme co-referent segments, the next task is to determine the theme of each segment. Using the Stanford Parser, every sentence in a theme co-referent segment is converted into standard Chomskeyan Tree Structure. It generates a constituency based parse tree. Every sentence in the tree is divided into Noun Phrase (NP) and/or Verb Phrase (VP). Sometimes, the sentence may also contain a Pronoun Phrase (PP). From the linguistic aspect, the Noun Phrases are indicative of the foci or the objects in the sentence—this is what is being discussed in the sentence, while the Verb Phrases denote the action between the objects. The noun phrases from every sentence in a segment are extracted to obtain the keywords or core topic words that describe the theme of the segment.

3.3 Segment-Wise Sentiment Analysis

For computing, the polarity for a segment S_i, the positive score count pos_sc_i and negative score count neg_sc_i of the segment are both initialized to zero. Then for every word in the segment, its label is checked to ascertain if it is positive or negative. For a positive word, the positive score count pos_sc_i is incremented by the score of the word. For a negative word, the negative score count neg_sc_i is incremented by the score of the word. Stop words are not removed for this task. At the end, the normalized positive score $normsc\_p_i$ is calculated as the ratio of positive to total scores. Similarly, the $normsc\_n_i$ represents the ratio of negative to total scores. If normalized positive score is greater than normalized negative score, segment is labeled positive, else it is labeled negative. The degree of polarity of the segment is denoted by $OrigPol_i$ which is a nonnegative array containing the normalized positive and negative sentiment scores, $normsc\_p_i$ and $normsc\_n_i$.

4 An Application in Sentiment Analysis

We describe an application in SA which can benefit from the segregation of a document into theme co-referent segments. On topic annotated data, the keywords of all the theme co-referent segments in a document compared to the overall document topic may indicate the more relevant segments and the overall document polarities adjusted to magnify or diminish the polarity scores of the segments as per their relevance to the topic being discussed in the document. This is helpful as we are able to segregate the relevant portions of text from supporting text such as analogies, anecdotes etc. For a document D on a topic T composed of a total of

k sub-themes {*S(1), S(2),…S(k)*}, the polarity weighing factor *w(i)* for a sub-theme *S(i)* having set of keywords *Keywords(i)* is computed by Eq. 1.

$$w(i) = \frac{\sum_{Key \in Keywords(i)} EGO(T, Key)}{\sum_{i=1\ to\ k} \sum_{Key \in Keywords(i)} EGO(T, Key)} \quad (1)$$

Thus, weight *w(i)* is the sum of the Extended Gloss Overlap (EGO) between the topic and all the keywords of a sub-theme divided by the Extended Gloss Overlap between the topic and all the keywords of each of the sub-themes. Extended Gloss Overlap, also known as Adapted Lesk, is a measure of Semantic Relatedness between two concepts as given by WordNet [16]. The revised polarity scores *Pol(i)* of each sub-theme *S(i)* are next computed by multiplying the weight obtained for the sub-theme by Eq. 1 to the original polarity scores *OrigPol(i)* for the segment as explained in Sect. 3.3.

$$Pol(i) = w(i) * OrigPol(i) \quad (2)$$

The final polarity scores for the document are calculated as the summation of the polarities of all the segments. This is represented mathematically by Eq. 3. In this equation the index j varies from 0 to 1 representing positive and negative polarities respectively and indicating that positive polarities are added to positive and negative to negative.

$$Polarity = \sum_{i=1\ to\ k} Pol(i)_j \quad , \forall j \quad (3)$$

5 Experiments and Results

We evaluated the proposed scheme on two sample documents containing context switches. They are shown in Fig. 1a, b. We explain our results and their significance with reference to these two samples. For the sample in Fig. 1a, the dendrogram generated after hierarchical clustering is as shown in Fig. 2. The x-axis indicates the sentence numbers in the document and the y-axis indicates the distance between sentences or clusters. It can be clearly seen from the dendrogram that a context switch occurs and the document was divided into two theme co-referent segments. Similarly, the document in Fig. 1b was also divided into two theme co-referent segments. On extracting the theme keywords from all the sentences in the two segments obtained in Fig. 1a, we found out that the first segment was about Sherlock, Code and Logic and the second segment was about Sir Arthur Conan Doyle, Sherlock, short stories, Psychological Disorder, Challenging and New Problem. For the document in Fig. 1b, the keywords obtained for the first segment were Goa, perfect potpourri, Portuguese charm, beautiful buildings, Fontahinas, Hindu Goan architecture, Nozomo, certain area, street food, Mandovi, Arabian,

(a)

I am in some ways like Sherlock. I love challenges. I am not so good at reading code and when I do not code hard enough, I find it a trouble to read my own code. In fact, when I do not code hard; my code is all code. When I code really hard, I find that I am reading my code effortlessly which I take as an indication of my code becoming less and less code. Logic flows flawlessly like a river when I code hard. Thus, the harder I code, the easier it gets to understand my code which I take as an indication of my code losing its code-like features. Sir Arthur never directly implied anything in the four novels or in the fifty six short stories. But one thing is clear: Sherlock was unable to deal with the consequence of not having a challenging problem to tackle. He would become so deeply agitated during these times that he would resort to the use of cocaine to fill his mind up with something. May be this sort of behavior can be classified as a mental or psychological disorder of some kind. But it is worth noting that he was never dependent upon cocaine in any way. He was not an addict. He would snap out of his usage almost instantaneously if he was presented with a new problem. Also, apart from the use of cocaine, he hardly had any other effects of not having anything to do with his mind.

(b)

Goa is the almost perfect potpourri of so many things. It has that Portuguese charm with all those beautiful buildings and houses (especially the Fontahinas area) and the Hindu Goan touch as well. It is comparatively less polluted, clean and beautiful. It has a centre for the arts (Kala Academy), NoMoZo (The roads of a certain area are closed for vehicles and people can use that road for anything, be it skating, dancing, quizzing, whatever), trekking groups, heritage walks and so many things. Goa has the best restaurants and street food, with varying cuisines. Goa is located at the mouth of river Mandovi and the Arabian Sea and that view at sunset is very beautiful. Though it's always the same river, same sea, same sun, it tells a new story every day. Thank you Quora for giving me my next travel destination. It will be the best location for my year end trip. Quora is seriously such a vast source of knowledge but sometimes it depresses me. It brings me a notch down because I can never compare to the knowledge a lot of people have. Quora makes me feel inferior as there are thousands of greater, accomplished and far more experienced people than me.

Fig. 1 **a**, **b** Two sample opinionated text documents for detecting context switches and performing sentiment analysis

new story, and year end trip. The keywords obtained for the second segment were Quora, travel destination, vast source. On performing sentiment analysis, the degrees of polarities of the two samples are as listed in Table 1. The method for computing the degrees of polarities of documents is same as that for segments illustrated in Sect. 3.3, only difference being for documents we consider the entire text and not just sentences belonging to a particular segment.

We discuss both these cases one by one. For the sample in Fig. 1a, the negative polarity is more than the positive polarity. This sample is about the topic code and while describing his love for coding the author presents an analogy to the very famous fictional character Sherlock Holmes, describes his habits, and outlines the similarities between him and Holmes. The writing style of the author in the first segment is very well indicative of his passion towards coding. In the second segment, while describing the habits of Holmes, the author lists quite a few negative

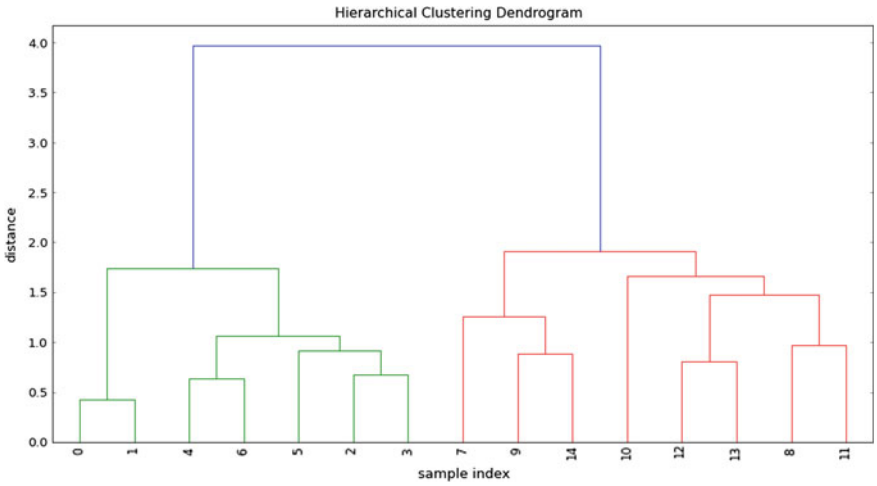


Fig. 2 Hierarchical clustering dendrogram obtained for theme co-referent segmentation of sample in Fig. 1a

Table 1 Sentiment polarities and labels of the samples in Fig. 1a, b

Sample	Positive polarity	Negative polarity	Label
1(a)	44.12	55.88	*Neg*
1(b)	64.06	35.94	*Pos*

connotation words and hence the negativity of the entire document increases. Though, his similarity to Holmes is not at all indicated in a negative way. As this document is about code, to classify it as negative because of the analogy used is incorrect. For the sample in Fig. 1b, the author discusses about Goa, a place he would love to explore. He goes on to state he decided upon that place by finding about it on Quora. And then there is a concept drift as the author from a positive mood and planning for his next holiday suddenly becomes dejected by the thought that how Quora, a storehouse of knowledge, is a constant reminder to him that there are numerous more well informed people than him. This document on the topic Goa is classified positive as the positive polarity is higher. But a quick examination can easily verify that the degrees of positive and negative polarities are far from correct. This is because the document is about Goa, and the author is only all praise for Goa.

In cases such as these, the proposed approach for tracking context switches and refinement of polarities thereafter is useful. We apply this scheme and divide both the documents into two segments and compute the segment-wise positive and negative polarities. Then, we compute the polarity weighing factors for the segments and using these compute the refined document polarities. These are presented in Table 2. It is worth noting that after adjustment of polarities, the sample in Fig. 1a, is now classified as positive. For the sample in Fig. 1b, the degree of positivity is refined to be more indicative of the degree of positivity of the topic being discussed in the document.

Table 2 Refined sentiment polarities and labels of the samples in Fig. 1a, b

Sample	Positive polarity	Negative polarity	Label
1(a)	58.02	41.98	*Pos*
1(b)	75.71	24.29	*Pos*

6 Conclusions and Future Work

Through this work, we proposed a scheme for SA which divides a text document into theme co-referent segments, thereby detecting the context switches present in text. We were able to verify that the proposed approach works well by testing it on two documents containing multiple themes. We further illustrated how this approach can be tapped to refine the accuracy of SA on text documents. We plan to improve upon this work by using an ontology to extract the themes conveyed by different segments of a text document. In future, we plan to test on a larger, real-world dataset.

References

1. Turney, P.: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the Association for Computational Linguistics pp. 417–424 (2002)
2. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In:Proceedings of the Conference on Empirical Methods in Natural Language Processing pp. 79–86 (2002)
3. Fiaidhi, J., Mohammed, O., Mohammed, S., Fong, S., Kim, T: Opinion Mining over Twitter Space: Classifying Tweets Programmatically using the R Approach. IEEE (2012)
4. Celikyilmaz, A., Tur, D.-H., Feng, J.: Probabilistic Model-Based Sentiment Analysis of Twitter Messages. IEEE (2010)
5. Shimada, K., Inoue, S., Maeda, H., Endo T.: Analyzing tourism information on twitter for a local city. In: First ACIS International Symposium on Software and Network Engineering (2011)
6. Jindal, N., Liu, B.: Mining Comparative Sentences and Relations, AAAI'06 (2006)
7. Ganapathibhotla, M., Liu, B.: Mining opinions in comparative sentences. In: Proceedings of the 22nd International Conference on Computational Linguistics pp. 241–248 (2008)
8. Hu, M., Liu, B.: Mining and Summarizing Customer Reviews. KDD'04 (2004)
9. Popescu, A.-M., Etzioni, O.: Extracting product features and opinions from reviews. In: EMNLP'05 (2005)
10. Mei, Q., Ling, X., Wondra, W., Su, H., Zhai, C.: Topic sentiment mixture: modeling facets and opinions in weblogs. WWW'07 (2007)
11. Nltk.stem Package. http://www.nltk.org/api/nltk.stem.html, Accessed 20 Jan 2016
12. Tf-idf. https://en.wikipedia.org/wiki/Tf%25E2%2580%2593idf, Accessed 20 Jan 2016
13. WordNet: A Lexical Database for English. http://wordnet.princeton.edu/, Accessed 23 Jan 2016
14. Lesk, M.: Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In: SIGDOC'86: Proceedings of the 5th Annual International Conference on Systems Documentation. pp. 24–26 (1986)

15. Esuli, A., Sebastiani, F.: Sentiwordnet: a publicly available lexical resource for opinion mining. In: Proceedings of the 5th Conference on Language Resources and Evaluation. pp. 417–422 (2006)
16. Banerjee, S., Pederson, T.: Extended gloss overlaps as a measure of semantic relatedness. In IJCAI'03 Proceedings of the 18th International Joint Conference on Artificial Intelligence. pp. 805–810 (2003)

Calendric Association Rule Mining from Time Series Database

Mudra C. Panchal and Ghanshyam I. Prajapati

Abstract In today's world, data explosion is high. Due to increase in Internet technologies and various applications, data are bursting. Out of numerous data, it is cumbersome to find out the interesting data and even interestingness of data differs from person to person, time to time, and task to task. Even the data keep on changing with time. Thus, an attempt is made to mine the important information from a large amount of time series data on a seasonal basis. An effort is carried out to mine calendric association rules, i.e., it will mine frequent itemsets based on the calendric pattern and generate association rules from it and the dataset considered is of time series dataset for market basket analysis on seasonal basis. FP-Growth algorithm is applied to carry out the task and the comparison is shown with respect to Temporal-Apriori and it is shown that FP-Growth is time efficient than Temporal-Apriori.

Keywords Association rules · Calendric association rules · Temporal apriori FP-growth

1 Introduction

Data mining has been widely researched now, but still proper research is needed for data that keep on changing with time, viz., temporal data. Many data mining task like association rule mining, classification, clustering, outlier detection, etc., are of utmost importance for many application. In this paper, focus is given on association rule mining. Up till now, much research has been carried out for association rule mining but less research for cyclic- or calendar-based association rule mining is done. Frequent patterns need to be found out from transaction database. Transaction database is temporal, viz., the time of a particular transaction carried out by customer is registered [1]. Here, the focus is to mine calendar-based association rules,

M. C. Panchal (✉) · G. I. Prajapati
Springer-Verlag, Computer Science Editorial, Tiergartenstr. 17, 69121 Heidelberg, Germany
e-mail: mudracpanchal@gmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_28

viz., the rules that occur for a specific instance of time. Calendric association rules are also called as seasonal rules as the frequent itemsets found do not occur throughout the database but only during some period of time [2]. Whereas cyclic association rules repeats itself at regular interval of time, viz., purchase of milk and bread daily during the time 9:00 A:M to 10:00 A:M [3]. Various algorithms are applied to mine such association rules which are described in detail in later sections but out of which FP-Growth algorithm [4] is more efficient in terms of execution time. Even an extension of Apriori, viz., Temporal Apriori is presented. It is specifically used for time series dataset [5]. A comparison between both the algorithms is also shown.

The paper is divided into sections. In Sect. 2, related work done by different authors is explained. In Sect. 3, preliminary terminologies are described. Section 4 consists of the problem statement of the work presented. In Sect. 4, proposed system is explained. Section 5 gives theoretical analysis for various algorithms. Section 6 consists of performance analysis and Sect. 7 shows conclusion.

2 Related Work

As per [6], author here explains two different algorithms, viz., sequential algorithm which consists of cycle pruning, cycle skipping, and cycle elimination, and interleaved algorithm. Results are shown by comparing the two algorithms based on dependence on minimum support, varying noise levels, varying itemset size, and data size scaleup. It minimizes the amount of wasted work done during the data mining process. Interleaved algorithm performs better than sequential algorithm. It does not provide updating of association rules. Numbers of frequent itemsets generated are more.

As per [5], author here has extended the famous Apriori algorithm to find temporal association rules based on calendar schema and calendar pattern. First, they identify two classes of temporal association rules, viz., temporal association rules w.r.t full match and temporal association rules w.r.t relaxed match. Then, they apply extended version of apriori named as temporal apriori that works level wise to develop two techniques that find association rules from both the classes of temporal association rules. It requires less prior knowledge. It also discovers more unexpected rules. Calendar-based temporal association rule mining can be done for other data mining tasks like clustering and classification. Time granules in a lower level must be obtained by subdivision of the time granules in higher level.

As per [7], author has proposed an efficient method which works differently from temporal apriori algorithm. This method scans the database at most twice. It works in three phases. First, it discovers frequent 2-itemset along with their 1-star candidate calendar patterns. In second phase, it generates candidate itemset along with their k-star candidate calendar patterns. In the last phase, it discovers frequent itemset along with their frequent calendar patterns. It avoids multiple scans of databases. It generates slightly more candidates than Temporal Apriori.

As per [8], author here has proposed an approach which consists of temporal H-mine algorithm and temporal FP tree algorithm. It also considers two parameters to mine the frequent itemset, viz., time and scalability. It is more efficient as it decreases processing time for mining frequent itemset. The approach is complex. It can be extended to design a good classifier.

As per [9], the author studied the problem of generating association rule that displays regular cyclic variation over time. Apriori algorithm is not efficient for such problem. Thus, the author has explained two new algorithms called sequential algorithm and transition algorithm. A new technique is devised called cycle rule by finding the relationship between association rules and time. The author has shown the difference between sequential and interleaved algorithm. Interleaved algorithm performs better than sequential algorithm. Implementing cycle rules reduces the amount of time needed to find association rules. Transition algorithm scales increasing data. It is not most efficient. Interleaved algorithm can be updated with minor changes in order to find global cycles.

As per [10], real time database keeps on updating and thus it is required to maintain and keep on updating the discovered temporal association rules. Implementing mining algorithm every time is inefficient as it does not maintain previously discovered rules and rescan the whole database. Thus, the author has proposed an incremental algorithm to maintain the discovered association rules. Results are shown based on both synthetic and real database. The proposed algorithm is ITARM which maintains temporal frequent itemsets after the temporal transaction database has been updated. The basic concept behind it is of sliding window filtering algorithm. It helps to reduce time to generate new candidates. It reduces rescanning of databases. It is scalable. It works efficiently only for maintaining the temporal association rules.

3 Preliminaries

In the previous section, we have seen types of work done by different authors and the strengths and weaknesses of them. Now in this section, a discussion of theoretical background required to carry out the proposed approach is presented. The different terminologies explained are support, confidence, dataset that is used, types of association rules, algorithms implemented for temporal association rules, etc.

3.1 Support

It is defined as a number of times a particular item or itemsets exists in a particular transaction database. It x and y are itemsets then, it is defined as the portion of the database that consists of both X and Y itemsets together. If the total records in

transaction database in 5 and only 1 records consist of both X and Y then the support is 20%. As per [1], the equation to find support is as follows:

$$\text{Support}(\text{XY}) = \text{Support count of}(\text{XY})/\text{Total number of transaction in D}$$

3.2 Confidence

It is defined as the portion of records that contains both X and Y to the total records that contain X. If the transaction table contains 10 records with X from total 20 records and confidence is 80% then 8 records out of 10 records that contains X also contains Y. As per [1], the equation defined is as follows:

$$\text{Confidence}(\text{X}|\text{Y}) = \text{Support } (\text{XY})/\text{Support } (\text{X})$$

3.3 Dataset

As the title says the dataset used for the implementation work is time series dataset, i.e., the dataset with timestamp, viz., either date or time or both. So the dataset used in the work is a food market dataset for carrying out market basket analysis. It is easily downloadable from the site: http://recsyswiki.com/wiki/Grocery_shopping_dataset [11]. It is a sample dataset from Microsoft and is available as mysql file.

3.4 Types of Association Rules

As per [12], various types of association rules are explained below.

Context-based Association Rules: Context-based association rules are classifies as a type of association rules which concentrate more on unseen (hidden) variable. These unseen variables are known as context variables which are responsible for changing the final set of association rules.

Generalized Association Rules: It is based on the concept of generalization where the hierarchy is climbed up. For example, city and state can be clubbed together to represent a state.

Quantitative Association Rules: It consists of both categorical and quantitative data [12]. For example, 70% of boys going to college will have a bike.

Interval Data Association Rules: Data are ordered and are separated by a specific range. For example, partition the salary of employees by 10,000 Rs. slots.

Sequential Association Rules: Association rules are ordered in sequence. For example, DNA sequence is important for gene classification.

Temporal Association Rules: Most of the real time data are temporal in nature. viz., it varies with time. Thus, we need to keep on updating the database. For example, purchase of AC by a customer in summer is seasonal; Purchase of milk in morning everyday is cyclic association rules and many more examples.

3.5 Types of Temporal Association Rules

Interval Association Rules: Data are ordered and are separated by a specific range. For example, partition the salary of employees by 10,000 Rs. slots. Each item is assigned a time interval so association rules are discovered during that time interval.

Sequential Association Rules: Association rules are ordered in sequence. For example, DNA sequence is important for gene classification.

Temporal Predicate Association Rules: A conjunction of binary temporal predicates is added to the association rule to extend it which specifies the relationships between the time stamps of transactions. It works for both point-based (Purchase of item at a fix time like sharp 10 o'clock) and interval-based (Purchase of item between a time interval like between 9:00 A:M and 10:00 A:M) mode.

Calendric Association Rules: It is based on calendar system. It is also called as seasonal association rules. For example, more accidents during rainy season, more purchase of refrigerator in summer.

Cyclic Association Rules: Ozden have introduced the concept of cyclic association rules and have shown the relationship between association rules and time [6]. For example, purchase of milk and butter daily during 9:00 A:M to 10:00 A:M.

3.6 Various Algorithms for Temporal Association Rules

Sequential Algorithm: It performs as per the fixed sequence from the starting to ending. Its work is to find association rules that are cyclic in nature or which repeats itself after certain period of time regularly, so it works in two steps. In the first step, association rules are found at a particular instance of time and in the second step, the cyclic patterns are detected. This method is implemented making use of Boolean expression that the association rule found is represented by true or 1 and if association rule does not exist then it is represented by false or 0. It is in the form of binary expression.

In the first step, whole binary expression is scanned and the one found with 0 is deleted and rest is saved to form association rules. This procedure continues till the end of last bit of binary expression.

In the second step, only large cycle is detected from the association rules found in first step.

Interleaved Algorithm: It is an extension of Apriori algorithm. It works in reverse to sequential algorithm. Interleaved algorithm first determines cycle or pattern that occurs regularly and then after, finds the association rules. It is more efficient than sequential algorithm. As per [13], it works in two steps but it discovers three more techniques.

Cycle omitting—It omits counting support for an itemset that does not belong to that particular cycle.

Cycle deletion—If the frequency of an itemset is less at a particular time, then it cannot have cycles. This allows omitting of cycles by cycle omitting step.

Cycle cutting—It simply prunes the cycles that are of no use.

As per [14], interleaved algorithm can be updated with minor changes in order to find global cycles.

PCAR Algorithm: It is better than both sequential and interleaved algorithm. This method works by performing the division between the original database into number of partitions or segments as per user wish. If there are 10 transaction and user wish to divide it into two segments, then first segment will consist of first five transactions and second segment will consist of last five segments. The segment will be scanned one after the other to generate the cyclic frequent itemset. After first segment, whatever cyclic frequent itemset is generated will be used to carry out scanning of next segment. Thus, it works in an incremental way. But the problem is it generates many association rules that are of no use to user.

CBCAR Algorithm: It is named as constraint-based cyclic association rules. It is an extension of PCAR. This algorithm eliminates the problem of PCAR of generating more number of association rules that are of no use to user. Here, user defined constraints will be followed by the rules. Thus, it minimizes the set of generated association rules.

IUPCAR Algorithm: It is incremental update PCAR. It is an extension of PCAR. It is more efficient than PCAR. As real time data are temporal so it is difficult to update them regularly as it requires rescanning of whole databases and hence IUPCAR came into existence as it solves the problem of rescanning of database.

It works in three steps. Three classes are made, viz., frequent cyclic itemset, frequent false-cyclic itemset, and rare cyclic itemset. Now in the first step, the database is scanned and the items are placed in its related classes out of three. In the second step, depending on the original class of the itemset and its support and new class and its support, an affectation of new class is made according to the weighted model [15]. In the final step, after the updating, final cyclic frequent itemset is found out.

FP-Growth Algorithm: As per [4], it is the algorithm that is used to mine the frequent itemsets from the dataset efficiently. It outperforms the working of Apriori algorithm and many others. It reduces scanning of database as it completes the whole procedure in just two scans of database. The main advantage of FP-Growth is it reduces generation of candidate sets. First is arranges the transaction is ordered sets and creates FP tree from the database and mines frequent itemsets directly from FP tree.

The algorithms described here can similarly also be used for other temporal association rules like temporal predicate association rules, etc. Along with these algorithms, fuzzy-based techniques can also be used to found out temporal association rules accurately.

4 Proposed System Description

A time series dataset for groceries is considered for the work. The dataset is available freely online with the attribute for time stamp as year, month, and quarter. Four quarters are considered with quarter 1 consisting of months 1, 2, 3, and quarter 2 for months 4, 5, 6 and quarter 3 for months 7, 8, 9 and quarter 4 for 10,11,12 months. Now as the work focuses on calendric association rules, a particular schema need to be considered. Thus here, the schema considered is (Year, Month, quarter) where the values are (1997, (1, 2, 3,…12), (Q1, Q2, Q3, Q4)). After defining a particular schema number of patterns possible is needed to be fetched. As per the dataset total, 12 calendric patterns are possible for 12 months. So the number of transactions for each pattern is generated. Now in order to find frequent itemsets for each pattern, 12 different fp-growth need to be applied, hence we have clubbed three patterns in one and applied fp-growth on it to find the frequent patterns for that particular pattern. The selected patterns are (1997, 1, Q1), (1997, 2, Q1) and (1997, 3, Q1) which is represented as frequent itemsets generated for all months in Quarter 1 in the year 1997, i.e., (1997, *, Q1).

Hence, the aim is to apply FP-Growth on the transactions for the pattern (1997, *, Q1) and generate all the frequent itemsets with its count and found calendric association rules for the pattern (1997, *, Q1).

4.1 Proposed System

The diagrammatic flow shown above is the overall approach for finding calendric association rules for time series dataset (Fig. 1). The proposed system makes use of FP-Growth algorithm which gives better results than Temporal Apriori which has already been used earlier. The execution time of Apriori Algorithm and its variations are more than FP-Growth algorithm. All the algorithms for mining association rules have been explained in the previous section and out of which Temporal

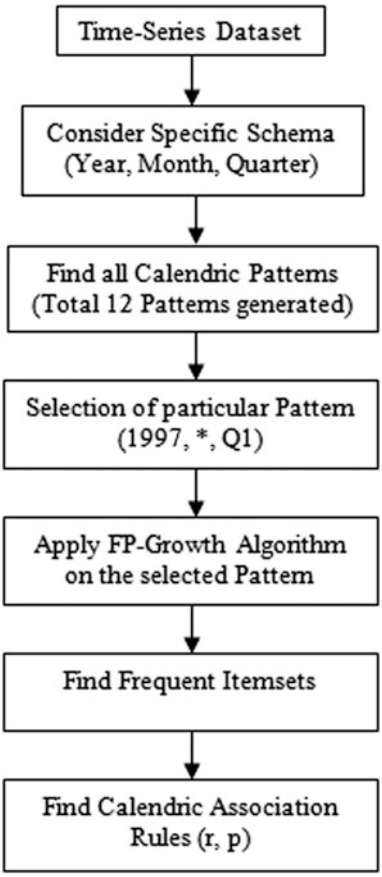


Fig. 1 Proposed system

Apriori works efficiently but the proposed system presented in this paper makes use of FP-Growth which executes faster than Temporal Apriori as it generates less candidates set and even scanning of dataset is only twice. The theoretical and practical analysis is described in later sections.

5 Theoretical Analysis

Below is the table representing the theoretical analysis done from the literature survey. Various algorithms have been studied and it has been analyzed that less work has been done in calendric association rule mining and out of all the algorithms FP-Growth is efficient to use (Table 1).

Table 1 Temporal association rule algorithms analysis

Algorithms	Strength	Weakness
Apriori	Simple to implement. It is widely used	Generate more candidate sets
Sequential	Extended from Apriori	Less efficient
Interleaved	Minimizes the amount of wasted work. More efficient than sequential	Generates candidate sets
PCAR	It outperforms sequential and interleaved algorithms	It generates rules that are not meeting expert's expectation
IUPCAR	Works fast for incremental update of cyclic association rules	Works on already generated cyclic association rules
T-Apriori	Generates less candidate sets.	Tedious work
FP-Growth	Generates less number of Candidate sets. Minimizes the number of scan of database	Requires more memory for storage of tree

6 Experimental Results and Analysis

Given below are the experimental results followed by the analysis of the algorithms presented here.

Experimental Results: The experiments are carried out to find the calendric association rules using the dataset of groceries [11] which is suitable for time series dataset. In Table 2, the experimental results for two different algorithms are shown, viz., Temporal Apriori and FP-Growth. The results are based on the execution time required for processing of dataset by both the algorithms. The execution time given is in milliseconds. The experiments are executed 10 times on same data and on same system. The experiments are carried out in java sdk 1.7 and Netbeans IDE 7.2 (Table 2).

Table 2 Temporal association rule algorithms analysis

Algorithms	Execution time in milliseconds	
	Temporal apriori algorithm	Fp-growth algorithm
Apriori	Simple to implement. It is widely used	Generate more candidate sets
Sequential	Extended from Apriori	Less efficient
Interleaved	Minimizes the amount of wasted work. More efficient than sequential	Generates candidate sets
PCAR	It outperforms sequential and interleaved algorithms	It generates rules that are not meeting expert's expectation
IUPCAR	Works fast for incremental update of cyclic association rules	Works on already generated cyclic association rules
Temporal Apriori	Generates less candidates sets	Tedious work
FP-growth	Generates less number of Candidate sets. Minimizes the number of scan of database	Requires more memory for storage of tree

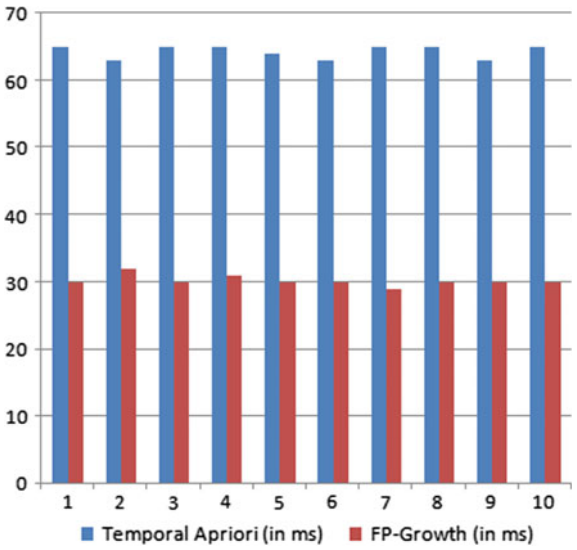


Fig. 2 Analysis chart

Experimental Analysis: From the Fig. 2, it has been observed that the execution time of FP-Growth algorithm is almost half than the Temporal Apriori algorithm. The experiments are executed 10 times and on an average, the execution time is much less in case of FP-Growth than Temporal Apriori. But the algorithms work differently for different datasets so the same algorithm may give different results on other datasets but the time taken by FP-Growth for execution will be always less than Temporal Apriori.

7 Conclusion

In this paper, a survey is presented on the various types of algorithms used for mining association rules. More focus is given on temporal association rules, i.e., that keeps on changes with time. Theoretical analysis is given for temporal association rule mining algorithm. The proposed approach explained here is based on FP-Growth algorithm which proves to work better than Temporal Apriori in terms of execution time. Comparison of various algorithms is given for the same. Special focus is given on time series dataset. Calendric association rules are mined from time series dataset. Still the performance of different techniques differs from each dataset as each dataset has their own characteristics.

References

1. Arora, J., Bhalla, N., Rao, S.: A review on association rule mining algorithms. Int. J. Innov. Res. Comput. Commun. Eng. **1**(5) (2013)
2. Shirsath, P.A., Verma, V.K.: A recent survey on incremental temporal association rule mining. Int. J. Innov. Technol. Explor. Eng. **3**(1) (2013)
3. Ale, J.M., Rossi, G.H.: An approach to discovering temporal association rules. ACM (2013)
4. Borgelt, C.: An implementation of fp-growth algorithm. IEEE (2011)
5. Li, Y., Ning, P., Wang, X.S., Jajodia, S.: Discovering calender-based temporal association rules. Elsevier-Data Knowl. Eng. **44**(2003), 193–218 (2003)
6. Ozden, B., Ramaswamy, S., Silberschatz, A.: Cyclic association rules. IEEE (1998)
7. Lee, W.-J., Jiang, J.-Y., Lee, S.-J.: An efficient algorithm to discover calendar-based temporal association rules. IEEE (2004)
8. Verma, K., Vyas, O.P.: Efficient calendar based temporal association rule. SIGMOD Rec. **34** (3) (2005)
9. Srinivasan, V., Aruna, M.: Mining association rules to discover calendar based temporal classification. IEEE (2008)
10. Gharib, T.F., Nassar, H., Taha, M., Abraham, A.: An efficient algorithm for incremental mining of temporal association rules. Elsevier (2010)
11. http://recsyswiki.com/wiki/Grocery_shopping_datasets. Accessed Oct 2015
12. Patel, Kaushal K.: A survey of cyclic association rules. IJEDR **3**(1), 453–458 (2015)
13. Shah, K., Panchal, M.: Evaluation on different approaches of cyclic association rules. IJRET **2** (2), 184–189 (2013)
14. Nanavati, N.R., Jinwala, D.C.: Privacy preservation approaches for global cycle detections for cyclic association rules in distributed databases, pp. 368–371. ResearchGate, July 2012
15. Ahmed, E.B.: Incremental update of cyclic association rules. Springer, Berlin, Heidelberg (2010)

Maintaining Bi-temporal Schema Versions in Temporal Data Warehouses

Anjana Gosain and Kriti Saroha

Abstract The temporal data warehouses (TDWs) have been proposed to correctly represent the revisions in dimension data. TDWs manage the evolution of schema and data with time through their versioning by time-stamping the data with valid time. Bi-temporal schema versioning in temporal data warehouses has also been discussed that not only allows retroactive and proactive schema revisions but also keeps track of them. The support for bi-temporal schema versioning raises an important issue for managing and storing the different versions of schema along with their data. The paper proposes an approach for managing bi-temporal schema versions in TDWs to allow for an effective management of several versions of schema along with their data.

Keywords Data warehouse · Temporal data warehouse · Schema versioning Transaction time · Valid time · Bi-temporal

1 Introduction

Data Warehouses (DWs) are a large reservoir of historical data, designed to provide support for multidimensional analysis and decision-making process [23, 25]. A DW schema as well as its data can undergo revisions with time to keep up with the application demands according to the user requirements and specifications. Several solutions have been presented in the literature to manage the revisions in DWs namely, schema and data evolution approach, schema versioning approach and temporal extensions. Schema and data evolution [5–7], [21, 22, 24, 28, 32] is a limited solution as it maintains only one DW schema and deletes the previous

A. Gosain (✉)
USICT, GGSIPU, Dwarka, India
e-mail: anjana_gosain@hotmail.com

K. Saroha
SOIT, CDAC, Noida, India
e-mail: kritisaroha@gmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_29

schema version incurring data loss. Schema versioning [2–4, 8, 9, 16, 27], on the other hand, preserves complete history of the DW evolution characterized by a set of schema revisions. But, it is also an established fact that not only the conceptual schema, but its underlying data may also evolve with time and thus, demand support for managing multiple versions of the schema as well as data. Temporal extensions [11, 13, 15, 26, 29], use time-stamps on the dimension data to fulfill this requirement and give rise to temporal data warehouse model with schema versioning support. TDWs use the research achievements of temporal databases and control the evolution of data with time-stamping of dimension data using their valid time. Valid time; determines the time for which an event is valid, and transaction time; indicates the time when an event is recorded in the database is generally used to keep a track and record of the revisions. At times, a combination of valid and transaction time (bi-temporal) may also be used [29].

Most of the works in the area of TDWs primarily deal with the evolution of dimension instances and use only the valid time for data and schema versioning. Schema and data versioning using valid time (Valid time versioning) is important for applications that require to handle retroactive (affecting things in the past) or proactive (what-if analysis) schema revisions but it fails to keep track of them (i.e., it only manages/implements the revision but does not keep track of when the revision was proposed). Bi-temporal versioning, on the other hand, not only manage but also keep track of retroactive and pro-active schema revisions, (i.e., keeps tracks of when a revision was proposed and when it was actually implemented) and has been discussed in the context of TDWs [18]. But, the paper does not discuss the storage options for the different bi-temporal versions of the schema as well as their data. This paper proposes an approach in the same direction.

In this paper, we aim to present bi-temporal schema versioning in TDWs with a wider aspect and discuss the storage options for several bi-temporal versions of schema and their data. The work extends the research achievements of temporal databases by time-stamping the schema versions with bi-temporal time [10, 19, 20, 30]. Two design solutions (central and distributed storage of data) are proposed to handle the extensional data when bi-temporal schema versioning is used. The support for bi-temporal schema versioning has been discussed at schema (intentional) along with data (extensional) level. Moreover, the concept of synchronous and non-synchronous alignment of data and schema is discussed in the context of TDWs (using valid time time-stamps on dimension data) and bi-temporal versioning of schema.

The rest of this paper is organized as follows. Section 2 presents a discussion on the related work. Section 3 gives an overview of different types of schema versioning and extensional data management. Section 4 presents the storage options for the management of bi-temporal versions of the schema with an example; and finally, in Sect. 5, we present the conclusions and final considerations.

2 Related Work

This section presents a discussion of the work done in the area of temporal data warehouses. Rasa et al. [1] recommended temporal star schema, using the valid time for time-stamping dimension and transaction data. Chamoni and Stock [11] proposed to store dimension structures with time-stamps containing valid times. Mendelzon and Vaisman [28] proposed a temporal design for multidimensional OLAP and also a temporal query language (TOLAP). The model apparently provided support for schema evolution by storing information related to the revisions in the instances and structure of the dimensions, but fail to record the history of data. A temporal multidimensional model was proposed by Body et al. [8, 9] that provides support to manage the evolution of multidimensional structures by using time-stamps on level instances along with their hierarchy and transaction data. However, only revisions to dimension schema and dimension instances have been discussed. Wrembel and Morzy [2] discussed Real and Alternate versions of multidimensional DWs but did not discuss the options to populate the new versions with data from previous versions. Golfarelli et al. [16] introduced the approach for schema augmentation but synchronous and non-synchronous mapping among data was not considered. The COMET model proposed by Eder and Koncilia [12–15] time stamps data with valid time to represent revisions in transaction and structure data. The model mainly deals with the evolution of dimension instances and does not include the evolution of schema or cubes. Mapping functions have also been proposed to allow transformations between structure versions using valid times but storage options have not been discussed for the versions. None of the approaches so far considered bi-temporal versioning of schema or data in TDWs.

3 Schema Versioning

Schema versioning done with respect to a single temporal dimension (transaction time or valid time) may be defined as transaction- or valid time versioning, respectively. Versioning that includes both temporal dimensions; transaction time as well as valid time produces bi-temporal versioning [14]. Thus, the versioning of schema can be categorized as transaction time, valid time or bi-temporal versioning.

- Transaction time Schema Versioning: Versioning of schema on transaction time, time-stamps all the versions of the schema with the related transaction time. It provides support only for on-time schema revisions, (i.e., revisions that are effective when applied) and that too, only in the current version of the schema. It does not allow for retro or proactive revisions.
- Valid time Schema Versioning: Schema versioning along valid time time-stamps all versions of schema with the associated valid time. The revised version of schema is effective only after its validity period is satisfied. In valid time schema versioning, multiple schema versions are accessible and any of the

schema versions can get affected by an update/revision if it either totally or partially overlaps with the valid time interval of the revision. It provides support for retro and proactive revisions but fails to keep a track of them.

- Bi-temporal Schema Versioning: This type of versioning time-stamps all the versions of schema using both transaction time and valid time. The transaction time is used to determine when the revision was suggested and the valid time indicates the duration for which the version of schema is valid. In bi-temporal schema versioning, only the present and the overlapped bi-temporal versions of schema can get affected by a schema revision. For a system that requires complete tractability, only bi-temporal schema versioning can establish that a new version of schema was generated as a result of a retro- or a pro-active schema revision [10].

Here, we propose design solutions for bi-temporal schema versioning in TDWs.

3.1 Design Choices for Handling Extensional Data

In temporal databases, two different storage solutions are discussed for managing extensional data namely, single pool and multi-pool solution [10]. We proposed to extend the concept for TDWs and presented two storage solutions [17]. The storage solutions and their response with respect to the bi-temporal schema versioning are discussed in this paper with the help of examples. The solutions are presented at the logical level, without moving into the physical design details.

Central Storage of Data, where only one data repository stores all data related to different versions of schema according to an extended schema, which contains all the attributes ever stated by any of the schema revisions.

Distributed Storage of Data, where multiple data repositories store data for various versions of schema. Each of the data repositories is configured corresponding to its related version of schema. To initialize a new storage, the records from the older storage are moved into the new storage according to the schema revisions.

In cases where both data and schema versioning are performed along the same temporal dimensions, any of the mappings; synchronous or non-synchronous may be used.

While using synchronous mapping, the version of schema having the same validity of records with reference to the common temporal dimension are used to record, extract and update data [10].

But in case of non-synchronous mapping, any of the versions may be opted to extract and update data irrespective of the valid time interval of the version of schema. Here, the validity of schema and data are independent even with regard to the common temporal dimension(s) [10].

Central storage of data is invariably synchronous, while distributed storage may be synchronous or non-synchronous.

4 Proposed Approach

Bi-temporal schema versioning preserves all the valid time versions of schema formed due to subsequent schema revisions. In bi-temporal schema versioning, only the current bi-temporal version of the schema that overlaps the validity interval specified for the schema update would be affected by the revisions. The user is allowed to select the bi-temporal schema versions to be included for modifications by specifying the validity of the schema revision. Moreover, bi-temporal schema versioning provides support for both implementing as well as tracking retro- and proactive schema alterations. The operations on bi-temporal schema versions and their data are described by means of examples and figures.

4.1 Managing Intensional Data

In bi-temporal schema versioning, a new version of schema is generated by implementing the required revisions to the present bi-temporal schema version that qualifies the specified valid time interval. The new bi-temporal schema version is assigned the validity period of the schema revision and current transaction time (NOW). Further, any of the older versions of schema that completely overlaps with the validity period of the revisions coexist with the new schema version but both have different transaction times. However, the validity interval of the older versions of schema, which have only partial overlap with the valid time interval of the revision applied is limited accordingly.

It is explained using an example as shown in Fig. 1. Suppose, an attribute p4 is removed from the bi-temporal schema version (SV) with validity [t′, t″]. The states of bi-temporal schema versions before the modification and following the modification are presented in Fig. 1a and b, respectively. Of all the bi-temporal versions of schema (SV_1 to SV_4) in the example, only two of them (SV_2 and SV_4) are partly overlapped and one (SV_3) totally overlaps with the valid time interval [t′, t″].

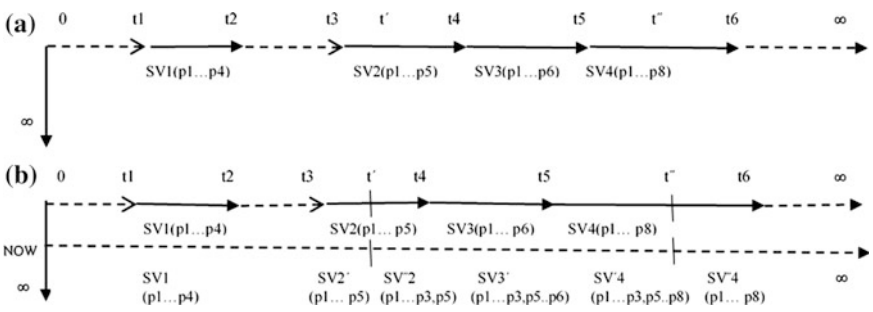


Fig. 1 Bitemporal versions of schema in temporal DW

The version of schema associated with [t_1, t_2] remains unaffected by the modification as it does not satisfy the valid time interval [t′, t″]. The schema versions associated with [t_3, t_4] and [t_5, t_6], partially overlaps the valid time interval [t′, t″] get affected by the modification and are thus split into two parts (SV′2, SV″2) and (SV′4, SV″4), respectively. The non-overlapping portion (SV′2, SV″4) is not affected by the modification and would retain all of its old attributes, whereas the overlapping portion (SV″2, SV′4) would remove the attribute p4 and results in (p1..p3, p5) and (p1..p3, p5..p8), respectively. The bi-temporal schema version related to [t_4, t_5] is also affected by the modification as it totally overlaps the valid time interval [t′, t″]. Therefore, p4 is removed from the new version of schema and creates SV′3 with attributes (p1..p3, p5..p6).

4.2 Managing Extensional Data

Since the TDWs mainly contains the valid time, bi-temporal schema versioning would be synchronous along transaction time and either synchronous or non-synchronous along valid time.

4.2.1 Central Storage of Data (Synchronous and Non-synchronous)

The central storage solution maintains only one data repository to store the complete data according to an extended schema version that contains all the attributes ever introduced by subsequent schema revisions [10]. The data repository can only grow, i.e., no attribute or temporal dimension is ever dropped from the data repository as a result of a schema revision and the revision can only be recorded in the meta-schema. But, if a schema revision results in the addition of an attribute, or a temporal dimension, the complete data repository is updated to the revised form. Data are thus stored using the enlarged schema format defined by successive schema revisions and the history of the revisions are recorded only in the meta-schema.

The information stored in meta-schema helps to restore the initial or previous structures of data for any of the bi-temporal versions of schema, if required.

4.2.2 Distributed Storage of Data

The distributed storage solution maintains different data repositories/stores for different versions of schema formed in response to schema revisions. A new data store is constructed according to the schema revisions and each version of the schema is allowed to access only its own data store. The new data store is populated with only the current data records from the older store associated with the modified version of schema and updated according to the revisions applied. The valid time

intervals of the records remain unaffected in the case of non-synchronous mapping, and their time-stamps are confined in the case of synchronous mapping.

Distributed Storage of Data (Synchronous).

Synchronous mapping of the distributed storage of data restricts the validity interval of data records in all the new data stores according to their intersection with the validity period of the version of schema. Also, for the management of data stores associated with versions of schema that have partial overlap with the revisions applied to the schema, the valid time interval of the records must be limited according to the valid time interval of their corresponding schema version.

Figure 2 shows the result of a modification (adding an attribute p3) which overlaps SV1 on [t2, t3], for synchronous mapping. It results in a new schema version SV2 that consist of attributes p1, p2, and p3. Figure 2a represents the initial data store SV1 with valid time interval [t1, t3]. The temporal pertinence and data records of distributed data stores are given in Fig. 2c. In the distributed storage solution, a new data store has to be created to support the new schema version SV2. The initial records are segmented in accordance with their valid time intervals and are subsequently divided between the two data stores. In the central storage solution, the records of the data store are not partitioned as shown in Fig. 2b.

It may be noted that the above constraint might result in loss of details about the original valid time interval of extensional data. It is evident from the example that for some records (e.g., (A1, B1)), the data may have duplicate copies in both newly created versions, where the validity period gets divided corresponding to the synchronous management of data.

Distributed Storage of Data (Non-synchronous).

Using non-synchronous distributed storage solution, the format of all data stores is according to the associated version of schema. Thus, a new data store is constructed and it is loaded with data by copying the data from the initial data store affected by the modifications specified for the older schema (update/drop of an attribute etc.).

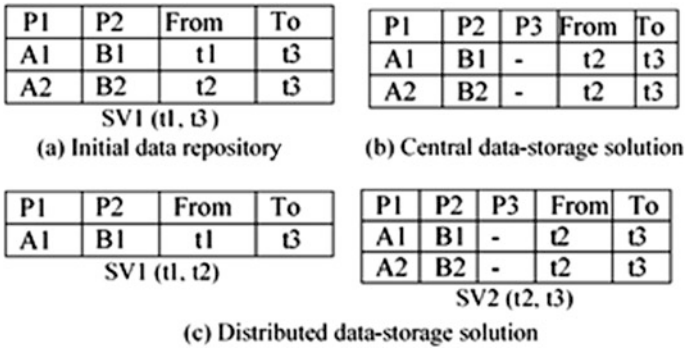


Fig. 2 Synchronous central and distributed storage solution

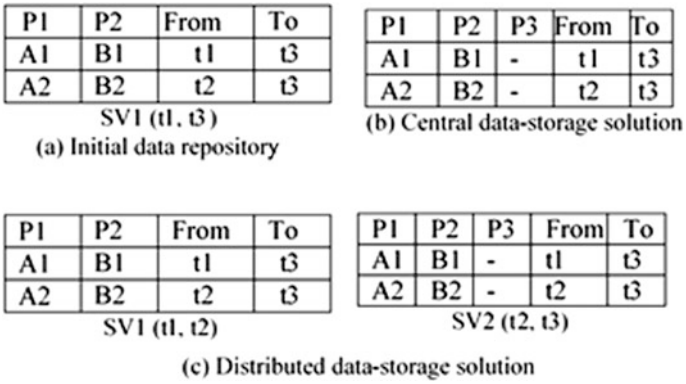


Fig. 3 Non-synchronous central and distributed storage solution

An example for non-synchronous mapping of extensional data is shown in Fig. 3. If the same schema modification is used (as given in Fig. 2) for the non-synchronous mapping, it does not require to partition the valid time interval for distributed storage of data. For this example, the results of central storage match with the new data store generated for distributed storage of data and it is notable that the data is not duplicated in this case.

5 Conclusion

The approach proposed in the paper would allow to trace the history of schema with bi-temporal versioning, assures consistency of data and presents different options to optimize the usage of space for storage of data. Bi-temporal versioning of schema and valid time versioning of data, together with the data storage solutions and the correlation between schema and data versioning, were analyzed to explore the possibilities to model a temporal data warehouse that supports bi-temporal schema versioning. The choice for the storage option is dependent on the availability of storage space; for example, if the available space is restricted then central storage of data may be selected. The preference for synchronous and non-synchronous management depends on the degree of freedom required among the data elements. The central storage solution does not create duplicate copies of data, but might require more storage space as it enlarges the format of the schema after revisions. On the other hand, the distributed storage duplicate data from the affected pool(s).

The distributed storage solution has an after-effect on the data model as the various data stores would support independent evolutions of the data by applying modifications through different versions of the schema. For some applications that need data from both older and newly created schema versions, it may be required to rebuild the entire history of schema versions if synchronously distributed storage is

used. This is because the records in the data stores are divided in synchronism with the distinct versions of the schema. On the other hand, in case of non-synchronous mapping, the full history of data is maintained for every version of the schema. Therefore it allows old applications to execute properly on older data as before, although newer applications are required to be developed for new data that contains the new attributes. Also, with synchronously distributed storage, data cannot be queried using the older details of schema because the data is updated only in the current data store.

Furthermore, if queries span over multiple schemas, when central storage solution is employed, then the solution generated for the query would comprise of only a single table. But, if distributed storage is adopted, then a new version of schema is required to be created that includes all the attributes needed for the solution of the query. The work can be extended to support bi-temporal schema versions in bi-temporal data warehouses.

References

1. Agrawal, R., Gupta, A., Sarawagi, S.: Modeling multidimensional databases. IBM Research Report, IBM Almaden Research Center (1995)
2. Bębel, B., Eder, J., Konicilia, C., Morzy, T., Wrembel, R.: Creation and management of versions in multiversion data warehouse. In: Proceedings of ACM Symposium on Applied Computing (SAC), pp. 717–723 (2004)
3. Bębel, B., Królikowski, Z., Wrembel, R.: Managing multiple real and simulation business scenarios by means of a multiversion data warehouse. In: Proceedings of International Conference on Business Information Systems (BIS). Lecture Notes in Informatics, pp. 102–113 (2006)
4. Bębel, B., Wrembel, R., Czejdo, B.: Storage structures for sharing data in multi-version data warehouse. In: Proceedings of Baltic Conference on Databases and Information Systems, pp. 218–231 (2004)
5. Benítez-Guerrero, E., Collet, C., Adiba, M.: The WHES approach to data warehouse evolution. Digit. J. e-Gnosis (2003). http://www.e-gnosis.udg.mx, ISSN No. 1665–5745
6. Blaschka, M., Sapia, C., Hofling, G.: On schema evolution in multidimensional databases. In: Proceedings of International Conference on Data Warehousing and Knowledge Discovery (DaWaK). Lecture Notes in Computer Science, vol. 1676, pp. 153–164 (1999)
7. Blaschka, M.: FIESTA: A framework for schema evolution in multidimensional information systems. In: 6th CAiSE Doctoral Consortium. Heidelberg (1999)
8. Body, M., Miquel, M., Bédard, Y., Tchounikine, A.: A multidimensional and multiversion structure for OLAP applications. In: Proceedings of ACM International Workshop on Data Warehousing and OLAP (DOLAP), pp. 1–6 (2002)
9. Body, M., Miquel, M., Bédard, Y., Tchounikine, A.: Handling evolutions in multidimensional structures. In: Proceedings of International Conference on Data Engineering (ICDE), pp. 581 (2003)
10. De Castro, C., Grandi, F., Scalas, M., R.: On Schema Versioning in Temporal Databases. In: Clifford, S., Tuzhilin, A. (eds.) Recent Advances in Temporal Databases, pp. 272–294. Springer, Zurich Switzerland (1995)
11. Chamoni, P., Stock, S.: Temporal structures in data warehousing. In: Proceedings of International Conference on Data Warehousing and Knowledge Discovery (DaWaK). Lecture Notes in Computer Science, vol. 1676, pp. 353–358 (1997)

12. Eder, J.: Evolution of dimension data in temporal data warehouses. Technical Report 11, Univ. of Klagenfurt, Dep. of Informatics-Systems (2000)
13. Eder, J., Koncilia, C.: Changes of dimension data in temporal data warehouses. In: Proceedings of International Conference on Data Warehousing and Knowledge Discovery (DaWaK). Lecture Notes in Computer Science, vol. 2114, pp. 284–293 (2001)
14. Eder, J., Koncilia, C., Morzy, T.: A model for a temporal data warehouse. In: Proceedings of the International OESSEO Conference. Rome Italy (2001)
15. Eder, J., Koncilia, C., Morzy, T.: The COMET metamodel for temporal data warehouses. In: Proceedings of Conference on Advanced Information Systems Engineering (CAiSE). Lecture Notes in Computer Science, vol. 2348, pp. 83–99 (2002)
16. Golfarelli, M., Lechtenbörger, J., Rizzi, S., Vossen, G.: Schema versioning in data warehouses. In: Proceedings of ER Workshops. Lecture Notes in Computer Science, vol. 3289, pp. 415–428 (2004)
17. Gosain, A., Saroha, K.: Storage structure for handling schema versions in temporal data warehouses. In: Accepted in 4th International Conference on Advanced Computing, Networking, and Informatics (ICACNI) (2016)
18. Gosain, A., Saroha, K.: Bi-temporal schema versioning in temporal data warehouses. In: Communicated in International Conference on Frontiers of Intelligent Computing: Theory and applications (FICTA) (2016)
19. Grandi, F., Mandreoli, F., Scalas, M.: A generalized modeling framework for schema versioning support. In: Australasian Database Conference, pp. 33–40 (2000)
20. Grandi, F., Mandreoli, F., Scalas, M.: A formal model for temporal schema versioning in object oriented databases. Technical report CSITE-014–98, CSITE-CNR (1998)
21. Hurtado, C.A., Mendelzon, A.O., Vaisman, A.A.: Maintaining data cubes under dimension updates. In: Proceedings of International Conference on Data Engineering (ICDE), pp. 346–355 (1999)
22. Hurtado, C.A., Mendelzon, A.O., Vaisman, A.A.: Updating OLAP dimensions. In: Proceedings of ACM International Workshop on Data Warehousing and OLAP (DOLAP), pp. 60–66 (1999)
23. Inmon, W.H.: Building the Data Warehouse. Wiley (1996)
24. Kaas, C.K., Pedersen, T.B., Rasmussen, B.D.: Schema evolution for stars and snowflakes. In: Proceedings of International Conference on Enterprise Information Systems (ICEIS), pp. 425–433 (2004)
25. Kimball, R., Ross, M.: The Data Warehouse Toolkit. Wiley (2002)
26. Letz, C., Henn, E.T., Vossen, G.: Consistency in data warehouse dimensions. In: Proceedings of International Database Engineering and Applications Symposium (IDEAS), pp. 224–232 (2002)
27. Malinowski, E., Zimanyi, E.: A conceptual solution for representing time in data warehouse dimensions. In: 3rd Asia-Pacific Conference on Conceptual Modelling, Hobart Australia, pp. 45–54 (2006)
28. Mendelzon, A.O., Vaisman, A.A.: Temporal queries in OLAP. In: Proceedings of International Conference on Very Large Data Bases (VLDB), pp. 242–253 (2000)
29. Schlesinger, L., Bauer, A., Lehner, W., Ediberidze, G., Gutzman, M.: Efficiently synchronizing multidimensional schema data. In: Proceedings of ACM International Workshop on Data Warehousing and OLAP (DOLAP), pp. 69–76 (2001)
30. Serna-Encinas, M.-T., Adiba, M.: Exploiting bitemporal schema versions for managing an historical medical data warehouse: a case study. In: Proceedings of the 6th Mexican International Conference on Computer Science (ENC'05), pp. 88–95. IEEE Computer Society (2005)
31 Vaisman, A., Mendelzon, A.: A temporal query language for OLAP: implementation and case study. In: Proceedings of Workshop on Data Bases and Programming Languages (DBPL). Lecture Notes in Computer Science, vol. 2397, pp. 78–96. Springer (2001)

Designing Natural Language Processing Systems with QuickScript as a Platform

Anirudh Khanna, Akshay, Akshay Garg and Akshita Bhalla

Abstract Chatbots are the most prolific examples of Natural Language Processing systems. These are computer software which can talk to users in natural language (via written and/or verbal methods). Like all other programs, these are also designed by computer programmers who are sometimes called "botmasters." There are some languages and development tools that are used to create chatbots, and QuickScript is one such platform. We started the development of QuickScript as an open-source platform for creating artificial conversational agents, with an intention to generate wider interest in the field of Natural Language Processing (NLP) and chatbot designing. The power of QuickScript is its simplicity and minimalism. QuickScript is relatively fundamental software with which one can design a chatbot in no time, without memorizing a lot of syntaxes and just paying attention to basic query-and-reply pairs. This paper will focus its discussion on introduction to the QuickScript project and using QuickScript for working with natural language systems.

Keywords QuickScript · Chatbot · Natural language processing
Artificial intelligence · Human–computer interaction

A. Khanna (✉) · Akshay · A. Garg · A. Bhalla
School of Computer Sciences, Chitkara University,
Chandigarh-Patiala National Highway (NH-64), Punjab 140401, India
e-mail: anirudhkhanna.cse@gmail.com

Akshay
e-mail: aakshay.740@gmail.com

A. Garg
e-mail: gargakshay.cse050@yahoo.in

A. Bhalla
e-mail: 27akshita@gmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_30

1 Introduction

Chatbots are becoming increasing popular and useful day by day, and practical chatbot applications are increasing with the spread of internet and the number of portable devices. Chatbots can be found in daily life, for instance as a source of entertainment, as help desk tools, to aid in education, for automatic telephone answering, and in e-commerce and business [1]. There are some languages and development tools that botmasters use to create chatbots, for example AIML (Artificial Intelligence Markup Language) [2] and ChatScript [3] are really efficient options these days for creating chatbots.

Our team developed QuickScript as such a platform which would facilitate the fast creation of simple chatbots with ease. The syntax of QuickScript is very simple and is composed of a minimal set of symbols that serve as keywords. Inspiration is taken from the language AIML for the features of QuickScript, but both the features and the way of writing chatbot code have been further simplified. It focuses on fundamental concepts that can be easily understood, like queries (any given user input, like a question or a compliment), patterns (a text which is to be matched with what the user has entered), responses (a set of replies that a chatbot will give when a pattern is matched), wildcards (for matching infinite number of words in between patterns), and so on. The following text gives a brief about the QuickScript platform, its basic elements and how to use it for starting a simple chatbot. The syntax is also given wherever required, along with appropriate examples.

2 The QuickScript Project

We developed QuickScript to set an example of effortlessness in chatbot design. It was started as an open-source project in January 2016 on GitHub. QuickScript has a special-purpose markup that is easy to learn. It can be used to make virtual conversational agents, programs that involve a dialog between the user and his/her devices. QuickScript can be employed to build interactive programs that may be something plain like a dictionary or something intelligent like a chatbot. Figure 1 shows a screenshot of QS Engine's user interface.

QuickScript is special because it is simple, coherent, and open source. Hence, it tries to fade the invisible barrier between a normal computer user and a botmaster (a person who creates chatbots). It is intended to generate a wider interest in the niche of AI and especially conversational agents, with introducing ease and speed in the design and development process. Here, it is vital to mention that QuickScript is inspired by the simple yet powerful development features of AIML (a markup language developed by Dr. Richard S. Wallace used by a number of Internet chatbots, including A.L.I.C.E.).

There are several ways in which AIML and the A.L.I.C.E. Bot defied the dominant computer science paradigms. According to Dr. Wallace, the ways of

Fig. 1 Home screen of QuickScript Engine v1.0

interaction in humans and computers are very different. Humans tend to spend a lot of time on real-time conversations and "chit chat"—informal dialog with little or no purpose. Computers, on the other hand, are known for giving precise and logical answers. Alicebot/AIML can be seen as a very effective attempt to bridge this divide [4].

QuickScript intends to establish the designing of conversational programs either in QuickScript itself or on any other platform like AIML as a popular and uncomplicated task among general people.

The project's website [5] and its GitHub repository [6] are complete resources for finding more information and downloading QuickScript files for local use (Fig. 2).

2.1 Behind the Idea

The idea of QuickScript came up during the development of a chatbot called S.A.R.A.N.G. [7] which was written in AIML, and here is when AIML influenced the development of the upcoming project by the set of features it offers precisely for the very purpose. We intended our project to be helpful in spreading the field of conversational agents to more and more people and so we decided to take the prominent features of AIML (wildcards, SRAI, etc.), simplify them, and make one of the easiest scripts for designing such entities.

Fig. 2 Webpage of QuickScript project hosted on GitHub [5]

2.2 Implementation, Software License, and Documentation

QuickScript code can be executed in the QS Engine (compiled as the "QuickScript.exe" file present in the project's main folder), which is written in C language (compiled on GCC Compiler 4.8). Being an open-source project, the source is available under the terms of GNU GPL v3.0 available online [8].

Complete QuickScript documentation can be found online [9] and not to mention many of the ideas and details included in this paper will easily relate to the QuickScript official documentation (needless to say, which is also prepared by the same authors).

3 Elements of QuickScript Code

The following section gives a brief introduction the various concepts and syntax elements seen in QuickScript programming. QuickScript documentation [9] can be referred for full details on QuickScript coding.

3.1 Entries

The code in QuickScript is basically lines of text that are called "entries." Each entry must be written in a separate line. It consists of two fragments—the "prefix" (set of symbols having specific meaning) and the "content" (whatever content the programmer decides the bot should recognize or learn or say).

3.2 Patterns

Any sequence of characters which the bot should be able to recognize is a "pattern." The knowledge base of a chatbot is all about the patterns it can recognize.

3.3 Responses

The chatbot's reply to a matched pattern is called a "response." For instance, "Hi! How is it going?" can be a suitable reply to the pattern "Hello."

3.4 SRAI

The term "SRAI" is borrowed from Artificial Intelligence Markup Language directly [10]. The part "AI" stands for Artificial Intelligence and "SR" usually stands for Synonym Resolution. SR can be different for different contexts; for example Syntactic Rewrite, Synonym Resolution, Symbolic Reduction. This reflects the multitude of ways in which SRAI can be useful. Basically, it can redirect a number of similar meaning queries towards a single answer, thus reducing the programming effort to a significant extent.

3.5 Wildcards

A wildcard character will match a number of words in a pattern where it occurs. QuickScript uses the asterisk (*) and underscore (_) symbols as wildcard characters.

3.6 External Learning

Currently, the External Learning feature is being developed, where the people who chat with the program can also teach it new replies to their queries. It can be enabled or disabled by the botmaster.

4 Working in QuickScript Environment: A Simple Program

Working with the current QuickScript implementation involves three easy steps:

- Writing code in a text editor
- Including that file into QS Engine
- Running QuickScript to see the results

4.1 Writing the Code in a Text Editor

QuickScript code can be written in any text editor. The file has to be saved with a ".qs" extension. Suppose a new file is created in the QuickScript folder with the name "stored_responses.qs," containing the following code:

```
> > HELLO
## Hello, user!
```

The two lines of code written above look quite simple and straightforward, yet they constitute a complete program in QuickScript, which has only one pattern and its one response. It is required to give a line feed (new line) between the first and the second line. Line feeds are required after each entry (line) in QuickScript.

4.2 Including the QuickScript Code (QS File)

Now, we need to include this file into the list of files that the QS Engine will interpret. This is a simpler step, but also as crucial as coding the file. For that, open "files.txt" present in the QuickScript directory. The list of the QS files to be run is mentioned in this file. Write the name of the newly created QuickScript file in "files.txt" in the following manner:

```
<path of file.../filename.qs>
```

For instance, if the name of the file is "stored_responses.qs" and it is in the same folder in which the QS Engine is placed, then including the file like this is sufficient:

```
< stored_responses.qs>
```

Once included, the file can be run by the QS Engine.

Fig. 3 Chatting interface of QuickScript

4.3 Running the Code in QS Engine

Start the QS Engine by running the file "QuickScript.exe," which is the actual program that will work with the QuickScript code written in the included file(s). Once the QuickScript interface is all set and the home screen comes up, load the included QS files by pressing ENTER and after that, the program enters the chatting interface, which means that bot is ready to chat with the user. The code should be interpreted without any errors if the files were written and included properly, and the program will prompt the user to enter a query. You can write a "Hello" and press ENTER. According to the code, it should print "Hello, user!" in response. A sample chat with a QuickScript bot is shown (Fig. 3).

5 Limitations and Scope of Development

QuickScript has been successfully implemented in various ways like in a simple dictionary program, a chatbot, a virtual doctor and so on. Admittedly, the project is relatively new and some issues like portability and online usage are required to be addressed. With some modifications in the underlying C implementation, these can be overcome. We hope to see the removal of any bugs in the platform and addition of newer developments in near future.

Acknowledgements Our sincere acknowledgement and thanks to all the people, resources, books and documents which helped in shaping of this work. Special thanks to ALICE A.I. Foundation and Dr. Richard S. Wallace for inspiration.

References

1. Shawar, B.A., Atwell, E.: Chatbots: are they really useful? LDV Forum, vol. 22, No. 1, pp. 1–2 (2007)
2. A.L.I.C.E. AI Foundation, Inc: AIML—The Artificial Intelligence Markup Language. http://www.alicebot.org/aiml.html (2016). Accessed 31 May 2016
3. ChatScript—sourceforge project web hosting: sourceforge.net: chatscript—project web hosting—Open source software. http://chatscript.sourceforge.net (2016). Accessed 31 May 2016
4. Wallace, R.: The elements of AIML style. Alice AI Foundation (2003)
5. Khanna, A.: QuickScript—easy to learn language for Artificial Intelligence. http://anirudhkhanna.github.io/QuickScript (2016). Accessed 5 June 2016
6. Khanna, A.: Project on GitHub, Inc. GitHub: anirudhkhanna/QuickScript. https://github.com/anirudhkhanna/QuickScript (2016). Accessed 5 June 2016
7. Khanna, A.: Pandorabots chatbot hosting platform: SARANG Bot—powered by pandorabots platform. http://pandorabots.com/pandora/talk?botid=9f0f09a71e34dcf8 (2016). Accessed 5 June 2016
8. Khanna, A.: Project on GitHub, Inc. GitHub: QuickScript/LICENSE. https://github.com/anirudhkhanna/QuickScript/blob/master/LICENSE (2016). Accessed 5 June 2016
9. Khanna, A.: Project on GitHub, Inc. GitHub: QuickScript/documentation. https://github.com/anirudhkhanna/QuickScript/tree/master/documentation (2016) Accessed 5 June 2016
10. Wallace, R.: The elements of AIML style. Alice AI Foundation, pp. 13–15 (2003)

Implementation of Low Cost, Reliable, and Advanced Control with Head Movement, Wheelchair for Physically Challenged People

Kunjan D. Shinde, Sayera Tarannum, T Veerabhadrappa, E Gagan and P Vinay Kumar

Abstract Hands and legs are the most important part of our mobility in day-to-day life. People will find difficulty in handling their daily activities if they have problems with their hands and legs (Physically challenged, accidental causes, and due to some health issues). Due to this incapability in movement causes several problems in their routine chores, and hence in order to provide a flexible mobility (stand-alone mobility to study, work, and day-to-day activities) in their life we came up this project "Low Cost, Reliable, Advance Control With Head Movements Wheel Chair for Physically Challenged People." In this project we are making use of Head Movements/tilts to control the Electronic Wheelchair for movements in all directions as per the need of the Physically challenged people, apart from head movements the wheelchair has certain propriety to another control signal (like Enable signal for Head control unit, Manual direction controls, emergency stop, power supply enable switch).

Keywords Head movements controlled wheelchair · Low cost wheelchair using arduino UNO · Head tilts using accelerometer · 25 A current driver for high torque motors · Distance sensing based head movement control
Reliable wheelchair for physically challenged people

1 Introduction

For a human being, mobility is the prime substance and need for several activities. Physically challenged people find difficulty in achieving mobility to do a given task, it is necessary for day-to-day activities like travel around, travel to work, and other activities.

K. D. Shinde (✉) · S. Tarannum · T. Veerabhadrappa · E. Gagan · P. Vinay Kumar
Department of Electronics and Communication Engineering, PESITM,
Shivamogga, India
e-mail: Kunjan18m@gmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_31

The need to facilitate mobility for physically challenged people drives us to design an embedded system which can provide the transportation and local mobility. Hence we came up with the project "Low Cost, Reliable, Advance Control With Head Movements Wheelchair For Physically Challenged People". This project is meant for physically challenged people (loss of limbs—legs/hands-and-legs—due to accidents, by birth, affected by certain diseases like polio, quadriplegia and so on). to facilitate a smooth and reliable form of mobility in their life to carry out their day to day activities.

1.1 Objectives

To design and implement an electronically controlled, real-time wheelchair for physically challenged people, where the Electronic control to the wheelchair is achieved by head movement of the person sitting on the wheelchair. The wheelchair is facilitated with safety features like automatic detection of the physically challenged person on the wheelchair, Emergency controls (priority based manual override, emergency stop, and system failure indication), and back drive distance monitoring as the salient features of the proposed wheelchair.

2 Literature Survey

The following are some papers and project work which we have referred for our work, In [1] the authors have designed and implemented head movement controlled a wheelchair using 8051 microcontrollers and the wheelchair carries a weight of 25–30 kg. (Complexity in the design is increased and accuracy is not achieved). In [2] the authors have implemented "Controlling an Automated Wheelchair via joystick Supported by Smart Driving Assistance", the control for a wheelchair is achieved by joystick and the wheelchair cost is high (Rs.60,000–65,000). In [3], the authors have implemented "Autonomous Wheelchair for Disabled people" In this additional assistance is required for wheelchair control. In [4] the authors have "Designing and Modeling of Voice controlled Wheelchair Incorporated with Home Automation", here the system need to senses the proper command and drive the wheelchair, voice modulation of individual and others have to be identified and the desired action has to be taken. (A. System is not reliable and for the new user, a new set of voice commands and recording has to be done. B. Implementation cost is high (Rs. 70,000/- to 75,000/-)). In [5] the authors have "Design and Development of a Hand Glove-Controlled Wheelchair Based on MEMS", control to the wheelchair is achieved by hand-glove control, this system is not effective for person with a disability in hands and legs. In [6–8] design and control of wheelchair using various other controls are mentioned and it working is coated.

3 Design and Working of Proposed Wheelchair

Figure 1 shows the block diagram of Head Control Unit, this is consists of Head tilt sensor which is used to sense the head movement motions and it is placed on a person's head. The analog signals from tilt sensor are given to Arduino UNO board and with the suitable programming the signals are calibrated and the control information is sent to base control unit via a wireless transmission unit interfaced at the head control unit. The power to head control unit is provided by a small 9 V DC battery. This unit is very compact and less weight so that it can be mapped to helmet/Cap as the physically challenged person can wear it.

Figure 2 shows the block diagram of the base control unit, here the signals from Head control unit are received from the wireless module, based on the security and wheelchair status the control signal to the motor and the wheelchair is driven as per the desire of the physically challenged person. Arduino UNO board dose does not provide a sufficient amount of current to motors so that the motors can carry the appropriate load and hence we are making use of a real-time high power motor driver to drive the motors. Additional safety features like Emergency stop and manual override is provided so that in case if head control unit malfunctions and to take a necessary action by others an alert sound in generated in case of such emergency. Will all the features, this wheelchair can be controlled by joystick/Keypad which is mounted on the wheelchair. For better control and home automation, we have internally designed a line following application for a wheelchair so that a local (home internal movements) mobility can be achieved.

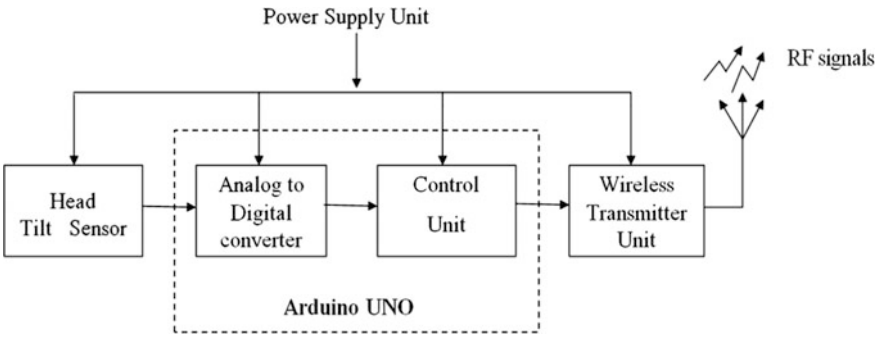


Fig. 1 Block diagram of head movement controlled wheelchair

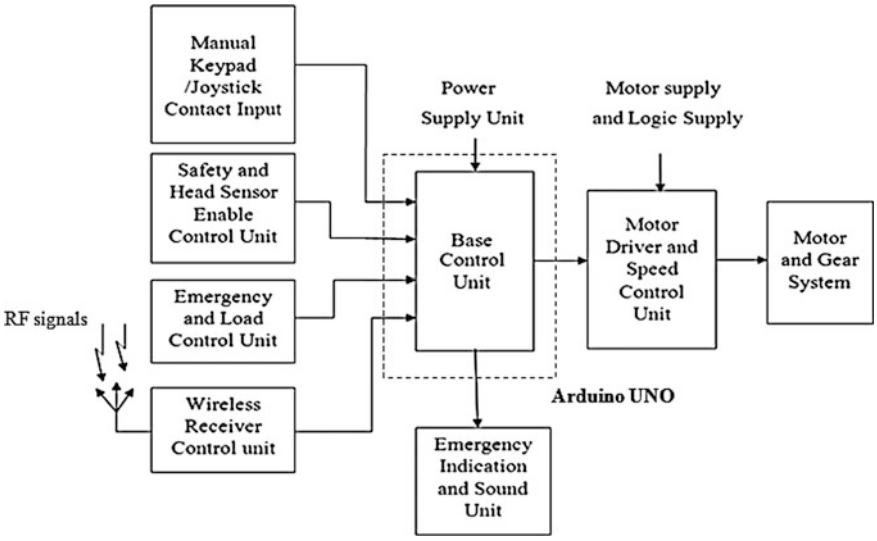


Fig. 2 Block diagram of base control unit for head movement controlled wheelchair

4 Results and Discussion

4.1 Design Implementation of Proposed Wheelchair System

The proposed system is implemented on the real wheelchair with high torque DC-geared motors which can drive up to a load of 25 kg and the entire module works for a real time. The Base control unit consists of RF receiver module, Distance sensing, 25 An H-Bridge current driver circuit, Control circuit with an emergency switch and Arduino UNO board to which all these modules are interfaced.

Figure 3 shows the implementation of the base control unit (BCU) and indicates the various modules placed on the wheelchair. The RF module is simple ASK Transmitter and Receiver module, the BCU consist of ASK receiver and other mounts of the wheelchair is indicated.

Figure 4 shows the implementation of the head control unit, it consists of MEMS accelerometer which is used to sense the head tilts and send the signal to BCU via ASK transmitter mounted on the head control unit. This system is powered by 9 V DC batter connected to the system.

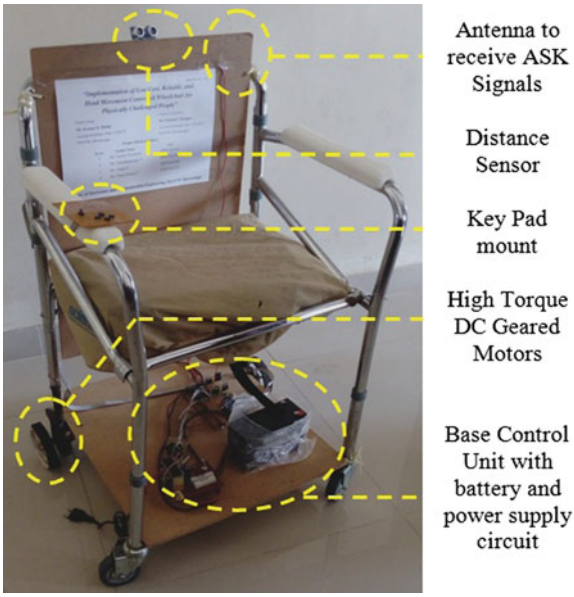


Fig. 3 Design of proposed wheelchair system

Fig. 4 Implementation of head control unit

The following figures show the images of various modules mounted on the wheelchair system so that the required output of the system can be obtained (Figs. 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17 and 18).

Fig. 5 Wheel chair for physical challenged people

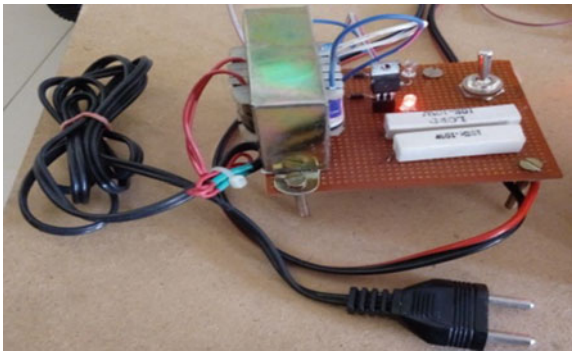


Fig. 6 Power supply system (to charge battery when it is LOW and to provide regulated supply and motor supply to the base control unit)

The above figure shows the implementation of the proposed wheelchair and from Table 1 we can analyze the behavior of the wheelchair on various inputs and its response to the state of the input applied with the priority assigned. It is clear that the emergency stop key is having the highest priority which is important to override any malfunction that may take place due to various problems.

Fig. 7 Lead acid battery for wheelchair system (12 V 7 Ah)

Fig. 8 Logic converter to H-Bridge motor driver (if the logic inputs to H-Bridge is logic '11' then the link is short circuited to avoid this we used the above shown converter which passes logic '00' when it gets logic '11' combination)

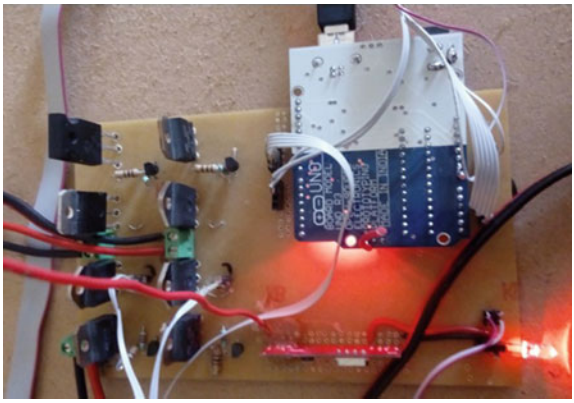


Fig. 9 Arduino UNO interface with motor driver and other units

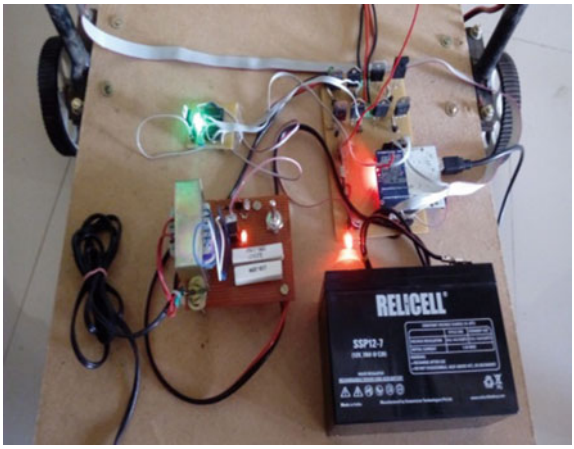


Fig. 10 Base control unit

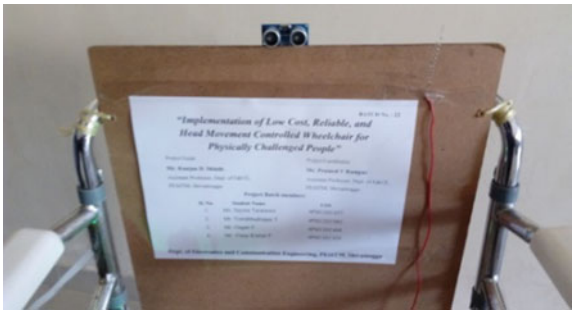


Fig. 11 Distance sensor and RF receiving antenna

4.2 Outcome of the Project

1. Motorized wheelchair is controlled and driven as per the head movements of the physically challenged person on the proposed chair.
2. Wheelchair is embedded with the salient features for security and reliability for the person on the wheelchair and works in real time.
3. Low cost implementation of the real-time wheelchair with the above facilities mentioned.

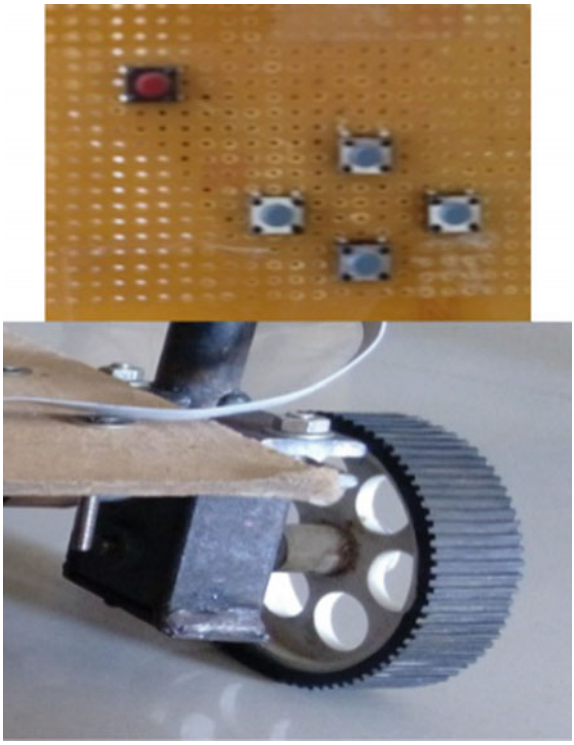


Fig. 12 Manual keypad and high torque DC motor

4.3 Applications

Electronically controlled Wheelchair can be used in

1. Hospitals (with modified control from head of the patient to hands of the Nurses in case of serious issue and hospital patient transportation)
2. General Home applications (can be used in office, home and college by the physically challenged person).
3. Local wheelchair can be replaced with the proposed wheelchair, which is very low cost and easy to use.

4.4 Advantages

1. Low cost compared to existing electronic wheelchair.
2. Head movements with joystick/keypad means of wheelchair control.

Fig. 13 Module testing for UP tilts

3. Motorized wheelchair with advanced control and safety features.
4. Reliable and cost-effective design.

5 Conclusion

The design and implementation of the wheelchair are done for real-time application, here the wheelchair works based on head movements when the head control is enabled and the keypad based operation is achieved based on the priority assigned to them. Hence the designed wheelchair is of low cost and reliable.

Fig. 14 Module testing for RIGHT tilts

Fig. 15 Module testing for LEFT tilts

Fig. 16 Enabling and disabling head movements using distance sensor

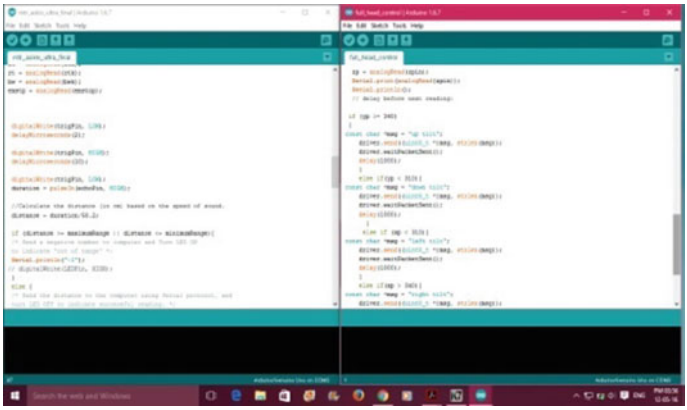


Fig. 17 Simulation of program using arduino software and serial monitoring the system

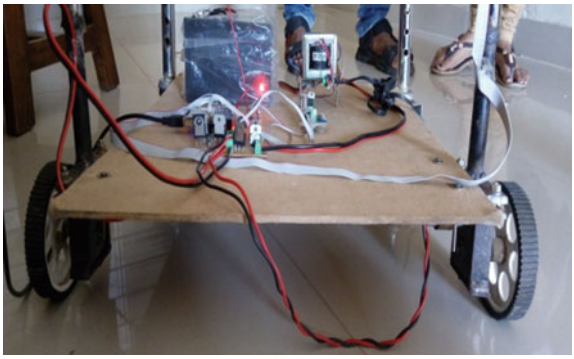


Fig. 18 Back view of the wheelchair base control circuit

Table 1 Function of wheelchair on the different actions of inputs

Distance sensor	< 10 cm	Head movements enabled	
		UP tilt	Wheelchair forward
		LEFT tilt	Wheelchair left turn
		RIGHT tilt	Wheelchair right turn
	> 10 cm	Head movements disabled	
Manual key pad and priority	Key	Priority	Wheelchair movements
	Emergency stop	1 (Highest)	Stand still
	Forward	2	Forward direction
	Left	3	Left direction
	Right	4	Right direction
	Distance sensor	5 (lowest)	Enable/Disable head tilts

Acknowledgements The Authors would like to thank the management and the Principal and the Dept. of E&CE, PESITM, Shivamogga, for providing all the resources and Support to carry out the project work. We would like to extend our heartfelt thanks to Ms. Tejaswini G.C, Asst. Professor, Dept. of E&CE, PESITM Shivamogga and Mr. Halaswamy K.E. Lab Instructor, Dept. of E&CE, PEISTM, Shivamogga for the help and motivation provided to carry out the project work.

Declaration

We the authors have obtained all ethical approvals from appropriate ethical committee and approval from the individuals for the study and the publication of accompanying images of participants in this study.

References

1. Kunjan, D.S., Raghuram, K.M.: Head Movement Controlled Wheelchair. A Project Works At Dept. of E&CE, SDMCET, Dharwad, June 2012
2. Rofer, T., Mandel, C., Laue, T.: Controlling an automated wheelchair via joystick/head-joystick supported by smart driving assistance. In: 2009 IEEE 11th International Conference on Rehabilitation Robotics Kyoto International Conference Center, Japan, 23–26 June 2009

3. Pires, G., Honório, N., Lopes, C., Nunes, U., Almeida, A.T.: Autonomous wheelchair for disabled people. In: Proceeding IEEE International Symposium on Industrial Electronics (ISIE97), pp. 797–801, Guimarães
4. Anoop, K.J., Inbaezhilan, Sathish raj, Rama seenivasan, Chola Pandian.: Designing and modeling of voice controlled wheel chair incorporated with home automation. Int. J. Adv. Res. Electr. Electron. Instrum. Eng. (An ISO 3297: 2007 Certified Organization). **3**(2) (2014)
5. Meeravali, S., Aparna, M.: Design and development of a hand-glove controlled wheel chair based on MEMS. Int. J. Eng. Trends Technol. (IJETT). **4**(8) (2013)
6. Tameemsultana, S., Saranya, N.K.: Implementation of head and finger movement based automatic wheel chair. Bonfring Int. J. Power Syst. Integr. Circ. **1** (2011)
7. Meshram, M.V.P., Rajurkar, M.P.A., Dhiraj Banewar.: Int. J. Adv. Res. Comput. Sci. Softw. Eng. **5**(1) (2015)
8. Puneet Dobhal, Rajesh Singh, Shubham Murari.: Smart wheelchair for physically handicapped people using tilt sensor and IEEE 802.15.4 standard protocol. In: Conference on Advances in Communication and Control Systems 2013 (CAC2S 2013) (2013)

Author Biographies

Mr. Kunjan D. Shinde is with PESITM Shivamogga, working as Assistant Professor in Dept. of Electronics and Communication Engineering and has a teaching experience of 2 Years. He is Pursuing PhD in VLSI-DSP domain, he received Masters Degree in Digital Electronics from SDMCET Dharwad in 2014 and received Bachelor Degree in Electronics & Communications Engineering from SDMCET Dharwad in 2012. He has published 13+ research papers in reputed journals like IEEE, Elsevier, Springer, and IOSR. His research interests include VLSI, DSP, Analog & Digital Electronics, and Robotics.

Ms. Sayera Tarannum is pursuing B.E. in Electronics and Communication Engineering from PES Institute of Technology and Management, Shivamogga, her area of interest include Embedded system design and microcontrollers.

Mr. T. Veerabhadrappa is pursuing B.E. in Electronics and Communication Engineering from PES Institute of Technology and Management, Shivamogga, his area of interest include Embedded system design and microcontrollers.

Mr. E. Gagan is pursuing B.E. in Electronics and Communication Engineering from PES Institute of Technology and Management, Shivamogga, his area of interest include Embedded system design and microcontrollers.

Mr. P. Vinay Kumar is pursuing B.E. in Electronics and Communication Engineering from PES Institute of Technology and Management, Shivamogga, his area of interest include Embedded system design and microcontrollers.

Part III
Computation Intelligence Algorithms, Applications, and Future Directions

Optimize Scale Independent Queries with Invariant Computation

S. Anuja, M. Monisha Devi and Radha Senthilkumar

Abstract Big data deals with the prodigiously and sizably voluminous volume of data engendered at high speed and it is arduous to process and manage with the subsisting database management tools. Query processing in astronomically immense data is a challenging task and frequently encounters problem. To overcome the complexity involved in processing the larger dataset, query optimization is the promising solution. Performance is a bottleneck, when complicated queries access an unbounded amount of data, resulting in high response time using the existing query optimization technique. In proposed work, to surmount this issue, scale independence is identified with access schema and query execution is optimized with invariant data. With astronomically immense precomputation and incremental computation dataset is used for querying. In precomputation approach, the invariant data is computed afore execution and thus resulting in lesser computation time during query processing. Incremental computation technique is applied to optimize the query for the streaming data. Thus, the invariant data is computed incrementally with the incipiently inserted data along with the precomputed data and then utilized for the query processing. By applying these approaches for optimizing scale independent queries, the performance can be ameliorated with tolerable response time.

Keywords Big data · Access schema · Query optimization · Precomputation Incremental computation

S. Anuja (✉) · M. Monisha Devi · R. Senthilkumar
SRM University, 5, Srinivasa perumal Koil St., Walajapet, Vellore 632513, India
e-mail: anujasubramani@gmail.com

M. Monisha Devi
e-mail: mmoishadevi.cse@gmail.com

R. Senthilkumar
e-mail: radhasenthilkumar@gmail.com

S. Anuja · M. Monisha Devi · R. Senthilkumar
Anna University, Chennai, India

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_32

1 Introduction

Big data contains a very large volume of heterogeneous data that is being generated at high speed. Therefore, current database management technology is not opportune for such data sizes and there is a clear desideratum for profoundly and immensely colossal data processing. In subsisting work, queries are optimized with the convergence property by elongating incremental evaluation and view materialization. In proposed work, the issues in optimization of queries are addressed with respect to scale independence which guarantees bounded amount of work for executing queries on a dataset independent of underlying data size. The proposed work highlights the consequentiality of precomputation for optimizing queries. The scale independent query is processed with the access schema formed and optimization is done by two approaches called precomputation and incremental computation along with the identification of invariant and variant data involved in the query execution.

2 Related Work

A view selection and maintenance system for scale independent queries are discussed in [1] which incorporates static techniques for analysis and also ascertain the engendered views do not affect the performance. Query answering with access restrictions and integrity constraints are extensively studied in [2]. A declarative language called Performance Insightful Query Language (PIQL) is proposed in [3] which computes upper bound over key/value store operations for rendering scale independence. Answering queries with minimal access patterns has been discussed in [4, 5]. PIQL in [6] provides stringent bounds on the number of I/O operations for any query and it is designed solely for astronomically immense scale data-intensive applications.

Rewriting queries utilizing circumscribed access patterns with integrity constraints has been discussed in [4]. An architecture for data storage which assures consistent scalability has been proposed in [7] granting developers to state application concrete requisite, capitalizes on utility computing in order to provide cost efficacious scaling. Hadoop framework is extended in [8] and queries are parallelized across nodes with the map-reduce framework.

An optimization framework has been proposed in [9] for SQL-like map-reduce queries. A powerful query language says MRQL (Map Reduce Query Language) has been focused in this paper and it captures the computation in declarative form to optimize it in a better way. An incremental evaluation algorithm is proposed in [10] for computing the view derived from the relations for deductive and relational database systems. Recursive delta-based computation has been extensively discussed in [11]. The incremental maintenance of view has been discussed in [12]. Data-intensive processing and tackling large-data problems have been extensively discussed in [13] and the distinct approach for those problems are highlighted.

3 Proposed Work

3.1 Query Optimization in Big Data

Query optimization in big data mainly concentrates on the possibility of resulting the subset from a large dataset in an efficient manner. Our work highlights the consequentiality of scale independence for immensely colossal data queries. Achieving scale independence requires the access schema for the dataset. Access schema is defined as the additional information available in the relation that clearly specifies which part of the data can be efficiently retrieved from the dataset for answering a query. Furthermore, the query which is made scale independent can be optimized by the approaches say precomputation and incremental computation with the invariant data identification.

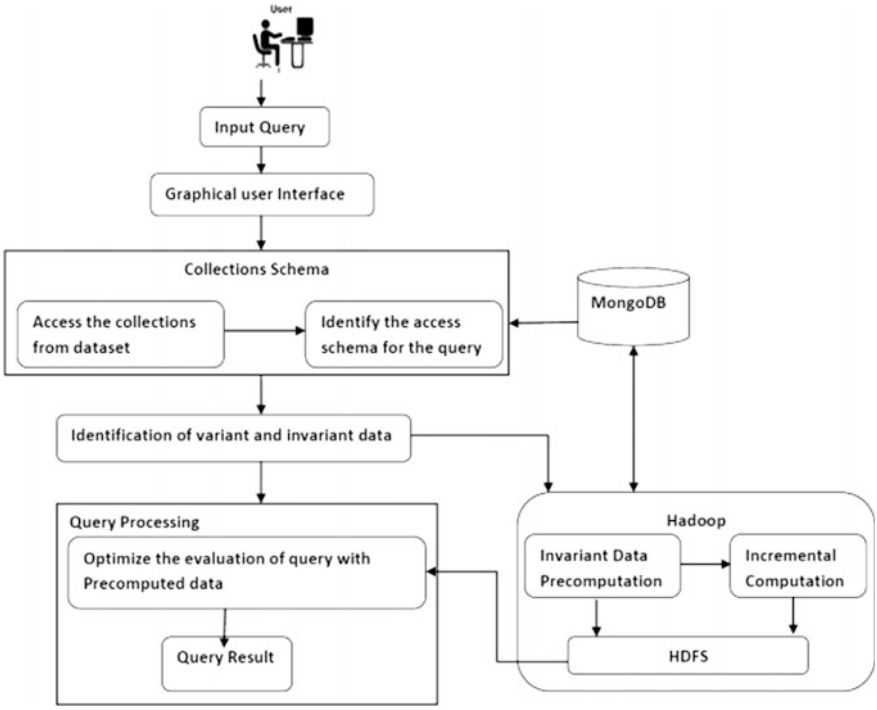


Fig. 1 Architecture of scale independent query optimization

3.2 System Architecture for Optimizing Scale Independent Query

The user posed query resembles the typical search of a signed in-user, in social networking sites like Facebook, twitter, etc. Fig 1 shows the architecture diagram for optimizing scale independent query in big data environment.

As per the architecture, when the user submits the query in the graphical user interface, the collections to be accessed for the query are identified from the dataset and access schema is analyzed for the user posed query based on the cardinality constraints existing between the collections of dataset. Access schema differs for every query according to the collections and relationship cardinality formed between the collections. Once the access schema is analyzed, then the variant and invariant data have to be identified with respect to the query based on the criteria which is applicable. After identifying the variant and invariant data, invariant data is precomputed and stored in HDFS. For further queries, these precomputed data is accessed instead of accessing the entire dataset. If there are updates in the collections, then the invariant data can be computed incrementally to answer the posed query. Scale independent query can be optimized with MongoDb a NoSQL database in Hadoop environment aiming at reducing the computation time. There are three modules in optimizing scale independent query and the modules are identification of invariant and variant data, invariant data precomputation, and incremental computation. Each of the modules is described in detail below.

3.3 Identification of Invariant Data

The identification of variant and invariant data is done through the normal execution process. Once the user submits the query, the associated collection from the dataset has to be identified. After identifying the associated collections, the access schema has to be analyzed based on the cardinality relationship formed between the collections.

Once the access schema is formed, the query has to be split into sub queries to process in a step by step manner and the sub queries are to be executed and stored in the array. Then the attribute values in the array which contains the intermediate result, as well as, the desired result are compared one by one within the array and categorized as variant and invariant attribute. Once the categorization is done, the invariant data which will retain during the entire querying process is computed priorly afore execution and stored in HDFS dynamically. The time taken for this execution process is displayed.

3.4 Invariant Data Precomputation

With the precomputation approach, the invariant data is computed and stored in HDFS priorly mainly to reduce the query execution time by reusing it in the querying process. Consequently, the query is executed as follows with precomputation. For the given query, the collections to be accessed are identified and access schema is predicated on the dataset involved in the query. Then the query is splitted into sub queries. After the split up, the sub queries are executed using the precomputed invariant data. Then the time taken for the query execution is exhibited along with the query result.

3.5 Algorithm for Invariant Data Precomputation

```
Input:D is the Dataset,Ci is the Collection in Dataset, Ai
is the Attribute, Q is the Query, Qij is the subquery
  Output:Qr is the Query Result

initialize D,Ci
setup VT,IT
for each query Qij in Q
      do
      split Qij =Qi1(Ci)+Qi2(Ci);
for each attribute $Ai in Qi1
      if $Ai has matches in Ci
      put ($Ai,Ai )in Ar
      /* Ar is the array where intermediate result is
      stored*/
      endif
endfor
for  each attribute $Ai in Ar,Qi2
      if $Ai has matches in Ci
put Ai in IT/* Invariant Table*/
else
put Ai in VT/* Variant Table*/
      endif
endfor
for each Query Qij in Q
      if Qi1(Ci)==IT && Qi2(Ci) ==VT
      then
            Q=IT + VT
      Set Qr=R(Q);
      endif
endfor
display Qr
```

3.6 Incremental Computation

The incremental computation handles the query execution with deference to updates as insertion into the dataset. The incremental computation reuses the invariant data computed priorly in the anterior module along with the incipiently inserted data stored in the delta table. First, the submitted query is splitted into sub queries. With reference to updates, the invariant data along with the incipiently inserted data is computed as the combination of invariant data calculated from original dataset afore updates and invariant data computed from the delta table and stored as a single table called invariant table in HDFS.

3.7 Algorithm for Incremental Computation

```
Input:D is the Dataset,Ci is the Collection in Dataset,Ai
is the Attribute,Qi is the Query,Qij is the subquery

Output:Qr is the Query Result

initialize  D,Ci, ΔCi,
initialize VT,IT
for each Qij in Q
      do
         split Qij =Qi1(Ci+ ΔCi)+Qi2 (Ci+ ΔCi);
            Qi1= IT+ Qi1(ΔCi)
            Qi2 =VT +Qi2 (ΔCi);
endfor
for each attribute $Ai in Qi1
if $Ai has matches in ΔCi
      put  ($Ai,Ai  )in  Ar/*  Ar  is  the  array  where
intermediate result is stored*/
endfor
for  each attribute $Ai in Ar,Qi2
if $Ai has matches in ΔCi
      put Ai  in IT
      else
      put Ai  in VT
endif
endfor
for each Query Qij in Q
         if (Qi1(Ci+ΔCi)==IT)&&VT =Qi2(Ci+ΔCi)
         then
            Qr=IT + VT
      endif
endfor
 display the  query result Qr
```

4 Experimental Analysis

With MongoDb, Hadoop has been acclimated for storing and processing astronomically immense volumes of data. Most of the approaches harness the potency of Hadoop and MongoDb together to process queries from a sizably voluminous data. MongoDb powers the authentic-time operational application, while Hadoop consumes all the data in MongoDb and coalesces it according to the query. All the collections of the dataset are loaded in HDFS which in turn stores a facsimile in MongoDb. The collections of the dataset are processed from MongoDb and the results are stored in HDFS.

Facebook dataset is utilized for the experimental purpose because relationship model and the constraints available in this dataset suit the project. Facebook.com is a convivial networking site and sanctions users to enter "friend" relationships with as many users, as one can accept with the constraints. Facebook dataset differs from other gregarious networking datasets like Twitter, LinkedIn, etc., with the presence of constraints involved in the cardinality relationship between relations.

4.1 Example

To find the friends of a particular person:

```
Schema: person (id, name, location)
friend (p, id)
friendlist(id,name,location)
Initialize
Setup friendslist;
Evaluate
Set friendslist=Select friend.p, person.id, person.name,
person.location from friend, person where person.pid=$p,
friend.id=person.id;
return query result;
```

The execution proceeds with the identification of invariant data in which collections are accessed to retrieve the desired result. Once the query is given, the associated collections of the dataset are to be identified and then the access schema is analyzed based on the cardinality relationship formed between the accessed collections. Then the query is split into sub queries and the sub queries are executed separately and the result stored in an array. Further, the attribute values in the array

which contains the intermediate result and the actual result are compared within the array. With the normal execution, the intermediate result and the desired result are compared. Then the person-id and friend-id are categorized as invariant for the given query, whereas the person name and location are subject to change and categorized as variant attributes.

4.2 Applying Invariant Data Precomputation

```
Schema: person (id, name,location
      friend(p,id)
IT(p,id)
VT(id,name)
friendlist(id,name,location)
initialize
      IT=select friend.p, friend.fid from friend where
friend.p=$p;
      VT=select id, name, location from person
evaluate
      Set VT=friendslist
      VT=Select VT.id, VT.name, VT.hometown from VT,IT
where  VT.id=IT.id
      return friendlist
```

As per the algorithm, the query is processed by applying invariant data precomputation. Instead of computing the entire query result right from the identification of collections, the precomputed data stored in HDFS has been reused. First, to predicate the invariant and variant data, the person relation, and friend relation is explored against the query. Then the invariant table is precomputed afore evaluating the given query. During the evaluation of the query, the query result is computed by utilizing invariant table and the variant table which gives the final query result. The invariant table which contains the friend-id of the given person-id is utilized to retrieve the name and location of the person from the person table by locating each person data with their id taken from invariant table. Consequently, the friend's name and location along with their id is determinately exhibited as the query result.

4.3 Applying Incremental Computation

```
Schema: person (id, name, hometown) and deltaperson (id,
name,hometown)
  IT(p,id)
  VT(id,name)
  friendlist (p,id,name)
  Initialize
      IT=IT + IT (ΔD)
        IT=select friend.p, friend.fid from friend where
      friend.p=$p and select friend.p,friend.fid from Δ
      friend where friend.p=$p
      VT =VT +VT (ΔD)
        VT=select id, name, hometown from person and
      select id,name,hometown from Δperson
  evaluate
      Set friendslist=VT
        VT=Select VT.id, VT.name, VT.hometown from VT,IT
      where VT.id=IT.id
  return VT
```

As per the algorithm, incremental computation has been applied as follows. Here the computation includes the incipiently inserted data in the dataset and query result is given by merging the invariant data precomputed from the old dataset and invariant data computed from incipiently inserted one. Consequently, the query result will contain the consummate friend list of the given person-id including the incipiently integrated friends and it is dynamically computed predicated on the insertion.

5 Performance Metrics

With Facebook dataset, the maximum number of instances for answering the query is considered and the execution time has been observed for different queries Table 1 represents the sample queries by applying the proposed approaches and their respective execution time. A comparison of execution time for precomputation, incremental with precomputation, and execution without precomputation approach has been made in the experimental analysis. With respect to Fig. 2, it is pellucid that the approaches result in tolerable response time.

Query 1 To find the complete list of friends of a particular person with their location.

Query 2 To find a person's complete friend list who likes a particular page liked by that person.

Query 3 To find the list of photos in a particular album in which a person is tagged by one of his/her friends.

Table 1 Performance metrics

Queries	Maximum number of instances to be accessed	Without precomputation (ms)	With precomputation (ms)	With incremental and precomputation (ms)
Query 1	5000	9668	669	218
Query 2	5000	7103	212	306
Query 3	5000	5234	2452	150

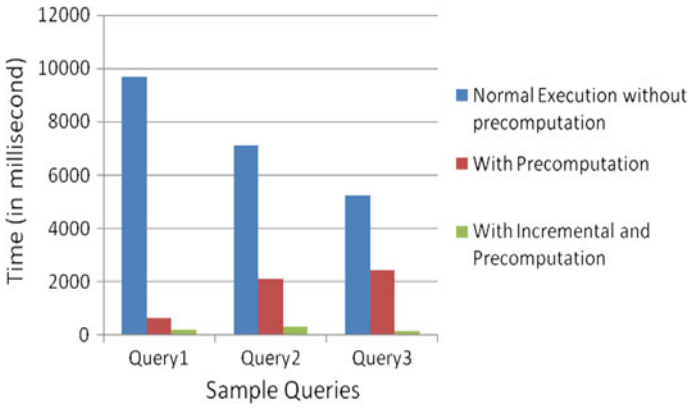


Fig. 2 Elapsed time for scale independent queries with MongoDb in Hadoop environment

With reverence to the proposed algorithms, the time taken for executing and retrieving the result of the query is drastically reduced, since it utilizes the invariant data precomputed through the identification of modifying and unmodifying attributes involved. However, for incremental computation, the time taken is lesser than precomputation since it utilizes the precomputed data along with the updates of dataset. The difference in the timing for these approaches limpidly designates that the queries are optimized in a better way by habituating these approaches with their respective algorithms. As per the experimental analysis with Facebook dataset, we have found that our algorithm gives a plausible response time with the proposed approaches.

6 Conclusion

With better utilization of access schema in query processing, the scale independence is made more facile. Along with scale independence, we have proposed two approaches such as invariant precomputation and incremental computation for optimizing the scale independent queries in Hadoop environment. An algorithm has

been proposed for both the approaches to optimize the queries with the identification of modifying and unmodifying attributes. With precomputation approach, the invariant data is utilized during the query execution for retrieving the result in lieu of accessing original dataset to reduce the time. The results are experimented with Facebook dataset by applying these approaches and a comparison is done between the ordinary execution, precomputation, and incremental computation. In future, we can experiment the query response time for grouping or aggregation of scale independent queries by utilizing map-reduce framework in a distributed environment and it can additionally be elongated to experiment with parameterized queries.

References

1. Armbust, M., Liang, E., Kraska, T., Fox, A., Franklin, M.J., Patteson, D.A.: Generalized scale independence through incremental precomputation. In: Proceedings of Special Interest Group on Management Of Data, pp. 625–636 (2013)
2. Barany, V., Benedikt, M., Bourhis, P.: Access patterns and integrity constraints revisited. In: Proceeding of the 16th International Conference on Database Theory, pp. 213–224 (2013)
3. Armbrust, M., Curtis, K., Kraska, T., Fox, A., Franklin, M.J., Patterson, D.A.: PIQL: success-tolerant query processing in the cloud. Proc. Very Large DataBase Endowment **5**(3), 181–192 (2011)
4. Deutsch, A., Ludascher, B., Nash, A.: Rewriting Queries using Views with Access Patterns under Integrity Constraints. Lecturer notes in Computer Science, vol. 3363, pp. 352–367 (2005)
5. Li, C.: Computing complete answers to queries in the presence of limited access patterns. Very Large DataBase J. **12**(3), 211–227 (2003)
6. Armbrust, M., Tu, S., Fox, A., Franklin, M.J., Patterson, D.A.: PIQL: a performance insightful query language. In: Proceedings of the International Conference on Management of Data, pp. 1207–1210 (2010)
7. Armbrust, M., Fox, A., Patterson, D.A., Lanham, N., Trusshkowsky, B., Trutna, J., Oh, H.: SCADS: scale independent storage for social computing applications. In: Proceedings of 4th Biennial Conference On Innovative Data Systems Research (2009)
8. Abouzeid, A., Bajda-Pawlikowski, K., Abadi, D., Silberschatz, A., Rasin, A.: HadoopDB: an architectural hybrid of MapReduce and DBMS technologies for analytical workloads. Proc. Very Large DataBase Endowment **2**(1), 922–933 (2009)
9. Fegaras, L., Li, C., Gupta, U.: An optimization framework for Map-Reduce queries. Proc. Extending Database Technol. 26–37 (2012)
10. Gupta, A., Mumick, I.S., Subrahmanian, V.S.: Maintaining views incrementally. In: Proceedings of Special Interest Group on Management of Data, pp. 157–166 (1993)
11. Mihaylov, S.R., Ives, Z.G., Guha, S.: REX: Recursive, delta-based data-centric computation. Publ. Very Large DataBase **5**(11), 1280–1291 (2012)
12. Koch, C.: Incremental query evaluation in a ring of databases. In: Proceedings of Principle of Database Systems, pp. 87–88 (2010)
13. Lin, J., Dyer, C.: Data Intensive Text Processing with MapReduce. Synthesis Lectures on Human Language Technologies, p. 177 (2010)

Generation of Optimized Robotic Assembly of Radial Engine

Rupalin Biswal and B. B. Choudhury

Abstract There are several procedures adopted for creating assembly sequence of a product. This research paper utilizes constrained method and LINGO method for creation of optimized robotic assembly. For creating sequences—constrained method is used. A modeling language which gives authorization to user to demonstrate model efficiently is developed for radial engine assembly. The result from two methods was evaluated and the best solution is suggested with a view of making the operation more economical.

Keywords Robotic assembly · Constrained method · LINGO

1 Introduction

Assembly is a procedure of forming the final product by putting a set of components parts together. Economy of a product is depending upon choice of assembly strategy. Assembly strategy depends on sequencing method. Finding effective sequence is very difficult because of two reasons. First, there are several feasible sequences for single product because of large amount of parts and second, minor changes in part design cause total change in assembly sequences. Also, an undesirable product is developed due to incorrect sequence which affects the total manufacturing process. An effective and good assembly sequence helps to reduce assembly cost. Nowadays with emerging technologies, it has become easier to analyze and estimate effective sequence for assembling product.

R. Biswal (✉) · B. B. Choudhury
Department of Mechanical Engineering, I.G.I.T, Sarang 759146, Odisha, India
e-mail: rupalin06@gmail.com

B. B. Choudhury
e-mail: bbcigit@gmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_33

2 Literature Review

Biswal et al. [1] estimated four different sequencing method of robotic assembly sequence. A number of predefined methods are considered and tried in randomly selected products. These products are used to develop different kinds of procedure to create robotic assembly sequences. Choudhary and Mishra [2] proposed two suitable techniques for development of assembly sequence in multi robotic assembly work cell. In above paper the sequence of robotic job developed by utilizing connecting graph and liaison method. Wang et al. [3] presented a three-dimensional model used for remote robotic assembly system. Robots are used to assemble the three-dimensional model using robot-mounted camera. This camera captures several pictures of the model at different poses. Generally, visions cameras and laser scanners are used to detect unidentified parts and transform those into three-dimensional models. Cai et al. [4] introduced that assembly modeling has a very significant role in analyzing product assembly. Assembly is a procedure of forming the final product by putting a set of components parts together. Dimensional control and quality improvement both are very difficult to analyze in variation analysis.

3 Outline of Work

The work in this paper is divided into two stages:

(i) Assembly of the product.
(ii) Optimization of assembly sequence.

Several assembly sequences obtained through constrained method. Thereafter, LINGO software is used on effective sequences to get optimized result for the given product. Paper is divided into several parts. Part II describes about the product taken for case study, constrained method, assembly and subassembly of the component, and set of algorithms. The flow diagram represents the steps of algorithm. After obtaining the assembly sequences it is used in LINGO that is given in Part III and Part IV. The result of Part IV is shown in images. Finally, Part V presents the conclusion.

4 Product and Procedure

4.1 Product

The product here is radial engine. Parts of radial engine are given below (Fig. 1).

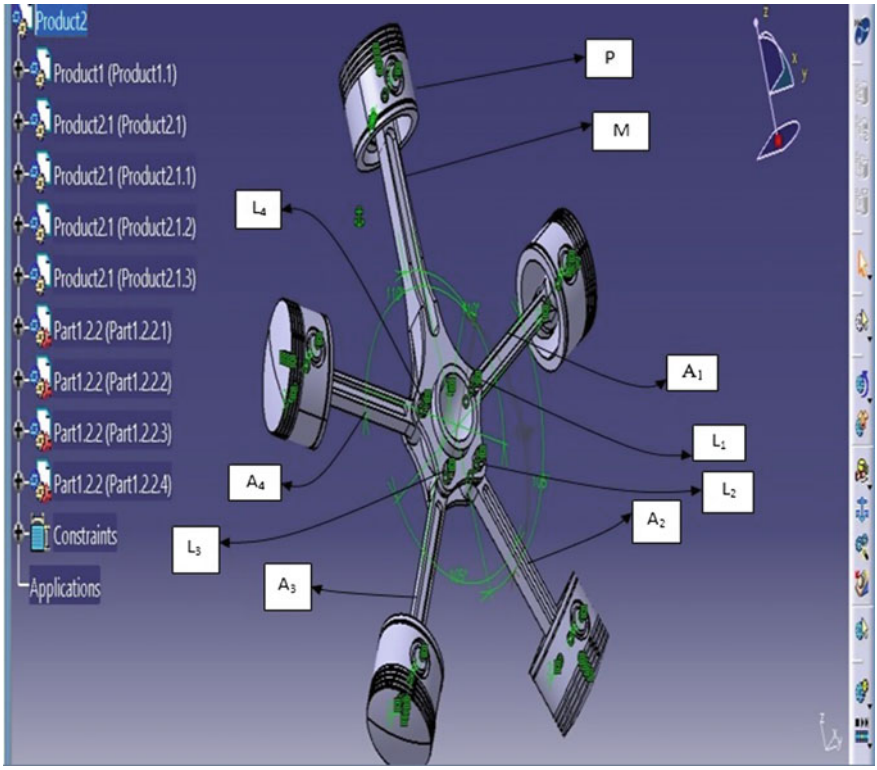


Fig. 1 Diagram of radial engine using CATIA V5

1. Master rod
2. Piston head
3. Articulate rod
4. Piston pin plug
5. Link pin
6. Piston ring
7. Bush

4.2 Procedure

As a result of various testing done by many scientists, several procedures have been developed for solving sequencing problem in assembly. The method study in this present work is constrained method.

Constrained method

Mainly two constraints are used in this method—G-constraint and C-constraint.

G-constraint:

1. One of the reasons is geometrical orientation of parts in the product.
2. During removal of a part, if it's unable to move in any direction due to some other part then the part has got 'G' constraint.

C-constraint:

'C' constraint occurs due to contact between different parts.

Basic procedure for constraint method

This process essentially gives the sequence of disassembling which is then used to generate the sequence of assembly of the product. The steps for obtaining the sequence in this method are given below:

Step 1 Analyze the given specimen.
Step 2 Then write down each part of the selected specimen and place them in one set called main set.
Step 3 Evaluate G-constraint(s) for every part in the product.
Step 4 Evaluate C-constraint(s) for every part.
Step 5 Components/subassembly(C/S) which do not have any constraints should be separated from the main set. At the beginning of the disassembly, it was removed to set an ordered manner. In some cases, there are multiple C/Ss which are free from constraints. In these types of cases separate the C/Ss in parallel manner and place it in different assembly set.
Step 6 Then reform the main set.
Step 7 Next step 6 and 7 are repeated until the main set becomes vacant.
Step 8 Then collect the parts separated sequentially to create disassembly order.
Step 9 Then check the final order and reject orders which are hard to create an assembly and are volatile in order to get some suitable sequences.
Step10 The reversal order of the disassemble pattern is called assemble sequence.

The total disassembling procedure is presented as a flow diagram, which may be modified and applied for searching single workable and balanced assemble orders (Fig. 2).

Abbreviation and Acronyms:

M Master rod
P Piston assembly
B Bush
L_1 Link pink 1
L_2 Link pin 2
L_3 Link pin 3
L_4 Link pin 4

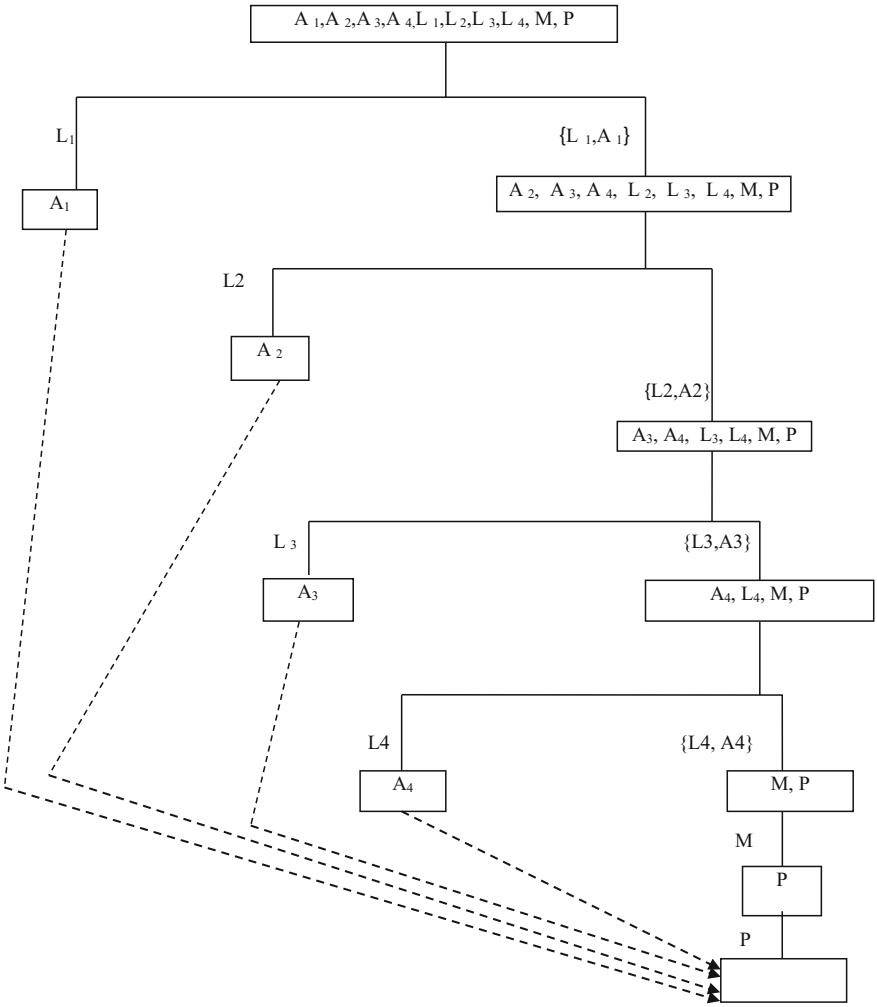


Fig. 2 Disassembly of radial engine when L1 taken first

P_H Piston head
P_p Piston pin plug
P_p' Piston pin
A_R Articulate rod
A Subassembly of articulate rod, piston head, piston pin plug, piston ring, piston pin, and bush. Since it has four subassemblies so I have assumed it as A_1, A_2, A_3, and A_4 (Fig. 3).

The effective assembly sequence of the above example is P-M-{L_1, A_1}-{L_2, A_2}-{L_3, A_3}-{L_4, A_4}. Similarly, 24 sequences are developed using the constrained method.

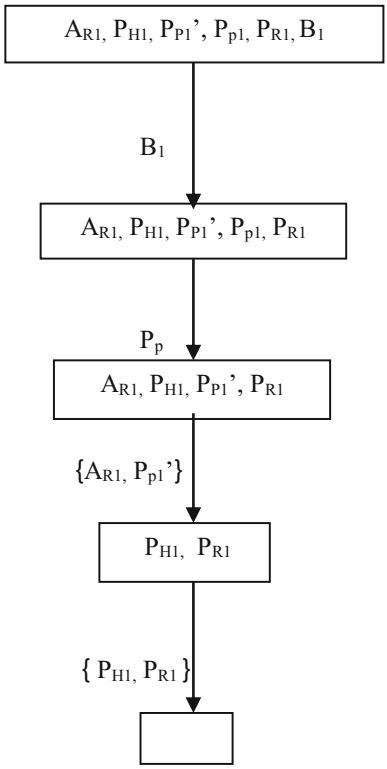


Fig. 3 Disassembly of subassembly A1

5 Lingo

LINGO is the most useful software due to its mathematical modeling language. LINGO software helps us to solve our problem in an easier way by using standard series of mathematical symbols. Instead of entering every item of each constraint in details, we can show a whole series of similar constraints in a single compose statement. Hence, it leads to models that are much easier to maintain and scale up.

6 Results and Discussions

The result is obtained by using constrained method for the product and to evaluate the assembly task it studied under LINGO 10. First, disassembly and assembly sequences are developed and then optimum order is obtained. Analysis and testing of model is done by applying these data.

6.1 Result of Constrained Method

Figure 4 shows the result which is obtained by using LINGO 10. Global optimal solution is found in LINGO Solver. Objective bound and Objective values have same result, i.e., 53. Integer Linear Programming (ILP) was adopted as a model class here. The total variables value is 45. Integer variable value is 44. Total constrained is found 26 and total nonzero value is 173. Infeasibilities, nonlinear variable, nonlinear constrains, nonlinear nonzeros results are zero. Final optimum sequence is found by adopting divergent disassembly and assembly orders in this method. The most effective assembly order is P-M-{L_1, A_1}-{L_2, A_2}-{L_3, A_3}-{L_4, A_4}.

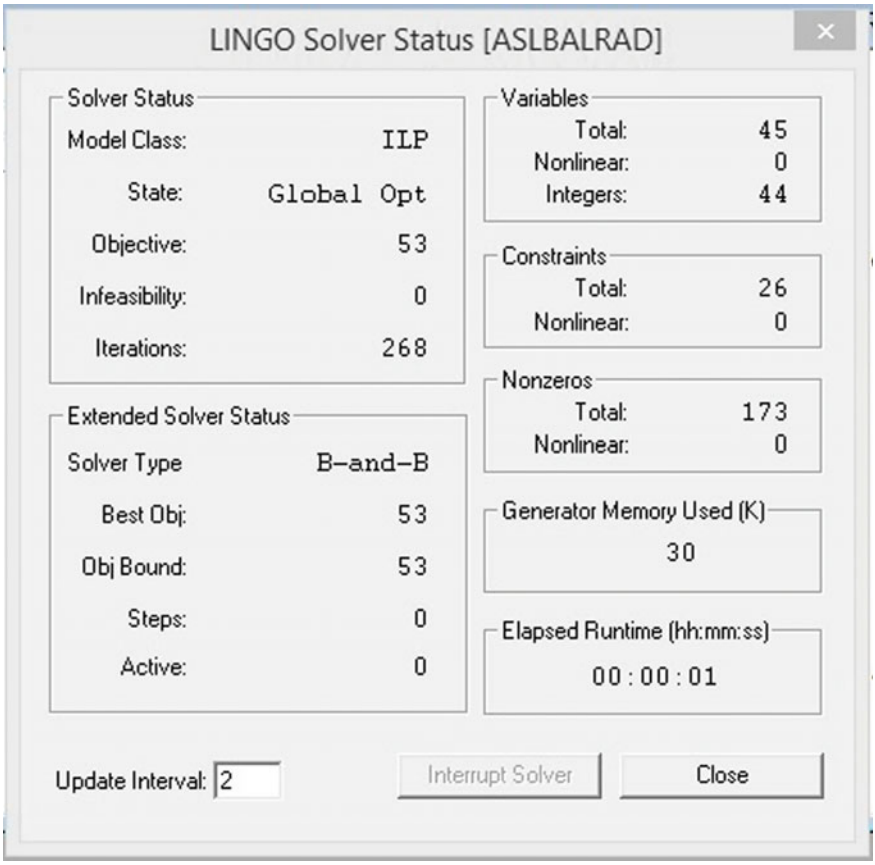


Fig. 4 Result box

7 Conclusions

The procedure adopted in this paper shows the importance on the number of assembly orders in which assembly task can be done. Our work in this paper is to create a suitable sequence for the robotic assembly of radial engine in less amount of time. Merits and demerits also exist in both methods. In case of product having larger number of components constrained method should take an account for assembly sequences. The existing experiment has given a different way for creating assembly orders which is used in robotic assembly process. Using programmable software for optimization it is easy to compare and obtain an effective sequence for assembly task.

References

1. Biswal, B.B., Choudhury, B.B., Mishra, D., Dash, P.: An overview and comparison of four sequence generating methods for robotic assembly. Int. J. Manuf. Technol. Manage. **20**(1–4) (2010)
2. Mishra, N., Choudhury, B.B.: Optimization of Sequences in Multi-Robotic Assembly Cells, AICTE Sponsored National Conference on Emerging Trend & its Application in Engineering (2011)
3. Wang, L., Mohammed, A., Onori, M.: Remote robotic assembly guided by 3D model linking to a real robot. Manuf. Technol. **63**(1–4) (2014)
4. Ciao, N., Qiao, L., Anwer, N.: Assembly model representation for variation analysis. Procedia CIRP **27,** 241–246 (2014)
5. Ortega, J.G., Garcia, J.G., Martinez, S.S., Gracia A.S.: Industrial assembly of parts with dimensional variation. Case study: assembling vehicle headlamps. Robot. Comput. Integr. Manuf. **27**,1001–1010 (2011)
6. Gonzalea, J.L., Rios-Cabrera, R., Ordaz-Hernandez, K.: On-line knowledge acquisition and enhancement in robotic assembly tasks. Robot. Comput. Integr. Manuf. **33**, 78–89 (2015)
7. Swain, A.k., Sen, D., Gurumoorthy, B.: Extended liaison as an interface between product and process model in assembly. Robot. Comput. Integr. Manuf. **30**, 527–545 (2014)

Velocity Restriction-Based Improvised Particle Swarm Optimization Algorithm

H. Mouna, M. S. Mukhil Azhagan, M. N. Radhika, V. Mekaladevi and M. Nirmala Devi

Abstract The Particle Swarm Optimization (PSO) Algorithm attempts on the use of an improved range for inertia weight, social, and cognitive factors utilizing the Pareto principle. The function exhibits better convergence and search efficiency than PSO algorithms that use conventional linearly varying or exponentially varying inertia weights. It also presents a technique to intelligently navigate the search space around the obtained optima and looks for better optima if available and continue converging with the new values using a velocity restriction factor based on the Pareto principle. The improvised algorithm searches the neighborhood of the global optima while maintaining frequent resets in the position of some particles in the form of a mutation based on its escape probability. The results have been compared and tabulated against popular PSO with conventional weights and it has been shown that the introduced PSO performs much better on various benchmark functions.

Keywords Swarm intelligence · Global optimization · Intelligent search
Inertia weight · Velocity restriction · Pareto principle

H. Mouna (✉) · M. S. Mukhil Azhagan · M. N. Radhika · V. Mekaladevi · M. Nirmala Devi
Department of Electronics and Communication Engineering,
Amrita University, Coimbatore, India
e-mail: mouna.harikumar@gmail.com

M. S. Mukhil Azhagan
e-mail: mukhil@outlook.com

M. N. Radhika
e-mail: radhikamnarayan@gmail.com

V. Mekaladevi
e-mail: v_mekaladevi@cb.amrita.edu

M. Nirmala Devi
e-mail: m_nirmala@cb.amrita.edu

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_34

1 Introduction

Particle Swarm Optimization (PSO) is a computational algorithm that simulates natural swarm behavior most notably found in certain species of birds, fish, and bees. Kennedy and Eberhart introduced this algorithm in 1995 [1, 2]. The particles in a swarm, analogous to flock of birds or fish in a school, move around the search space in a predefined pattern that is close to the natural search of a bird for its food which is the global optima. The communication between members of a flock is also simulated and each particle in the swarm is aware of its own optima and the best optima among the swarm. Different particles explore in different velocities and from different positions, all of which are randomized to obtain results close to the natural order of swarming [3].

PSO being faster in terms of iterations and processing time has progressed rapidly in recent years and has been used for problems in Artificial intelligence, Material design, and various fields that require a quick optimization [4, 5]. Conventional PSO gets stuck at a local minimum [6]. Research has been focused on trying to accelerate the convergence speed and achieving a better accuracy [7–9]. Reported literature has shown either a linearly decreasing pattern alone or a combination of two different algorithms, popularly known as hybrid algorithms, which are comparatively slow on convergence [10–12]. The improvised algorithm attempted in this paper has a varying inertia weight based on the Pareto principle, thus enabling it to converge to a better value with more precision.

To achieve the above-mentioned superiority over the other algorithms, three techniques have been introduced. The first technique, the *velocity restriction* is based on the Pareto principle (or the 80-20 rule) [16, 17]. The second technique is on a mutation-based term called the *escape probability*, which allows for a way of doing extensive exploration which focuses on finding other local optima easily. The third technique uses an improved *inertia weight* [10], which will progressively converge the search toward the minima over higher iterations. The designer is allowed to set the cognitive and social parameters, so that the user has control over the neighbor of convergence, either toward the pbest or gbest depending on the requirement of the algorithm. At every iteration, there is a condition for mutation that is based on its *escape probability*, which is the number of times it moves out of the boundary. This serves a dual purpose of maintaining the position within the search space of interest and also mutates the position of the particle frequently. This frequency is controlled by the escape probability, which is in turn controlled by the initial velocity that is set by the designer. Combining the three techniques—*mutation*, *velocity restriction*, and the refined *inertia weight* (based on the Pareto principle) the algorithm has been made adaptive and intelligent to work with varied functions. Simulation results have been produced to show better convergence and precision.

2 Particle Swarm Optimization

The algorithm was first proposed by considering a swarm of particles of size N [1, 2]. The optimum position in the search space was found by these particles using their swarm intelligence. The conventional PSO uses five basic principles namely Quality principle, Proximity principle, Stability principle, Diverse response principle, and Adaptability principle [1]. At the start of the PSO, the number of variables D is initialized and the objective *function f* is specified. The required parameters such as population size, swarm size, total number of iterations *itmax*, cognitive factor $c1$, social factor $c2$, and inertia weight w are initialized. The independent variables are given their boundary conditions in which they can search for the best optimum position. The random values of position and velocity are initialized as the pbest values for each particle, and gbest value for the swarm. *pbest* is defined as the personal best of each particle and *gbest* is the global best of swarm. In the given search space, the new position value is found by each of the particles. The position and velocity are calculated using the Eqs. (1) and (2) given below.

$$v_i^D = w * v_{i-1}^D + c1 * rand1_i^d * (pbest_i^d - x_i^d) + c2 * rand2_i^d * (gbest^d - x_i^d) \quad (1)$$

$$x_i^d = x_{i-1}^d + v_i^d \quad (2)$$

Where v_i^D is the velocity of the current iteration for each argument, w is the inertia weight [10]. $rand1_i^d$ and $rand2_i^d$ are two random distributions that range between 0 and 1. x_i^d is the position of each particle that will be updated from its previous value. The new position value is compared and checked with its preceding value. If the new position value is found to be better than the preceding value, the pbest value is updated else the preceding value is retained. The best among the *pbest* of all the particles is taken, if that value is better than the preceding *gbest*, then it is replaced as the *gbest* value, else the older value is retained. Inertia weight significantly affects the accommodation between exploitation and exploration in the PSO process. Different variants of PSO can be obtained by changing parameters such as the cognitive and social factor, different inertia weights, swarm size, network topologies in PSO, etc. [13]. Hybridization and multi-objective are some of the variants. In hybridization, for example, Genetic Algorithm and PSO can be combined; GAs mutation technique can be combined with PSO to prevent PSO from getting stuck at local optima [14].

3 Improvised PSO

The difference between conventional PSO [1, 2] and the improvised PSO is that it has techniques for *mutation*, *velocityrestriction*, and *improvised ranges* for factors that affect PSO. This PSO works just like the conventional PSO except for the fact that it has a restricted search in different range of weights, and velocity restriction based on the Pareto principle. The Pareto principle states that 20% of the cause is responsible

for 80% of the outcome or vice versa, depending on the perspective. Applying this, the *inertia weight*, which is calculated using Eq. (3) is varied from values between 0.8 to 0 to cover 80% of the area which is a higher neighborhood of exploration, to find the rest 20% of the local optima after convergence has started.

$$w_i = wmax - \frac{(wmax - wmin)}{itmax} * it \tag{3}$$

Where w_i is the weight of each iteration, w_max is around 0.8 and w_min is around 0. Values taken in the trials are 0.7 and 0.1, respectively. *itmax* denotes the maximum number of iterations considered and *it* denotes the current iteration. Also, a technique of velocity restriction, calculated using the Eq. (4) that modifies the effect of the preceding velocity on the existing position, is included.

$$V_r = e^{\frac{-it}{k*itmax}} \tag{4}$$

Where Vr is the *velocity restriction* factor. *k* is a constant, whose value is taken as 4 for an optimum range. Equation (4) should be multiplied along with velocity during every iteration to restrict its boundary. It decreases exponentially and the speed can be modified by changing the value of *k* by the user depending on the need. Frequent mutations are performed to help the exploration process, determined by an escape probability. This escape probability is calculated using an algorithm that also prevents the particle to move out of the boundary, thus stabilizing the swarm search. The Pseudo code for the improvised PSO is given in Fig. 1.

PSEUDOCODE

Step a: Initialize the number of dimensions, all the required parameters (the size of population and swarm, total number of iterations *itmax*, cognitive and social factors *and inertia weights)*, and specify the objective function.
Step b: Calculate the *inertia weight* for each iteration using equation (3), within the range 0.8 - 0.0.
Step c: Calculate the *velocity restriction* parameter for each iteration using equation (4).
Step d: Define the boundary conditions where the particles search for optimum position.
Step e: Initialize random values for position and velocity for particles as *pbest* and *gbest*.
Step f: Find the next *gbest* and *pbest* values.
Step g:Find the new position value of the particle
Step h: Find the new velocity of each particle using *velocity restriction* technique using equation (1) and equation (4).
Step i: Each individual position should be updated using equation (2) and the new velocity.
Step j: Restrict each particle to the defined boundary using the mutation technique, utilizing *escape probability*.
Step k: Compare and check the new position value with the preceding value. If the new position value obtained is better than the preceding position value, go to *m*, else go to *n*.
Step l: Update the preceding values of *pbest* and *gbest* with the new best value obtained.
Step m: Go to *Step o*.
Step n: Retain the previous values of *pbest* and *gbest*.
Step o: If i<=*itmax*, i++, go to step *g* else go to *step p*.
Step p: Indicate the optimum value of *pbest* and *gbest*.

Fig. 1 Pseudo code for improvised PSO

4 Simulation Results and Analysis

The improvisation performed on Particle Swarm Optimization (PSO) was tested on benchmark functions and the results acquired are improved comparatively [10–12]. The Code was simulated in Matlab using a PC with core I3 4005u, 1.7GHz, and 4 GB RAM. Benchmark functions used are (a) Sphere function, which is continuous, unimodal, and has D local minima (b) Rastrigin function, which is multimodal and has several local minima (c) Rosenbrock function, also known as valley or banana function, is unimodal (d) Michalewicz function, which is multimodal and has D local minima and are usually referred as valleys and ridges (e) Shubert function, which has many local and global minimas. The results have been obtained for 220 iterations and over 30 independent trials with 100 particles. The search has been done over a search space of $[-5.12, 5.12]$ for (a), (b), (c), and (e) functions and $[-\pi, \pi]$ for (d), as given in [18]. Table 1 shows the 2D plot between mean function value of all particles and number of iterations of Sphere, Rastrigin, Rosenbrock, Michalewicz, and Shubert functions, respectively, including equations and 3D plots.

Table 1 shows plots for the benchmark functions, for (a), (b), (c) the y-axis for the 2D plot is indexed in powers of 10, as the expected values from mathematical calculation [18] are close to zero. For (d), (e) the y-axis is in real values. Table 2 shows the iteration at which the minimum value is obtained for each benchmark function. Combining 2D plots from Tables 1 and 2, various inferences can be made. In Table 1 for the Shubert function, which is multimodal, a large number of spikes can be seen, which denote various particles exploring other parts of the search space for potential global minima until the 220 iterations end, whereas the minima has been reached at around the 26th iteration in Table 2. This demonstrates the ability of the algorithm to explore exhaustively even after getting settled at the minima, due to the mutation factor introduced using *escape probability*. For the sphere function, which has a single minima, the algorithm tries to obtain the best possible value, from Table 3, it can be seen that the algorithm is precise up to 10^{-71} on an average. In such functions, the algorithm is able to choose exploitation over exploration thus leading to much better values as proven in Table 4.

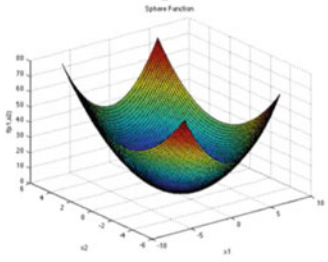
In Table 3, $X1$ denotes the position in the first dimension. $X2$ denotes the position in the second dimension and *fgbest* is the fitness value that is dependent on $X1$ and $X2$ as described by equations in Table 1. Table 3 shows the values of $X1$, $X2$, and *fgbest* using the Eqs. (5), (6), (7), (8), and (9) for various benchmark functions.

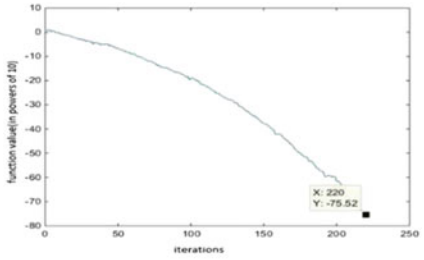
In Table 3, the best-fit column denotes the best value obtained. This value is the global minimum that has been obtained through the algorithm over the trials. The mean and standard deviation denote the algorithms variation from the best-fit value. It has been shown that the algorithm performs with minimal variance for most of the benchmark functions when the fgbest value is concerned. Incase of $X1$ and $X2$, the average and the standard deviation are close to the expected values, except in case of (e), the Shubert function, as the function exhibits the same minimum value at multiple points in the given search space. This discrepancy is expected out of such a function and can be verified mathematically [18].

Table 4 shows the fgbest values using the introduced algorithm, i.e., the fitness value and it is compared with [11, 12, 15]. The proposed algorithm exhibits much

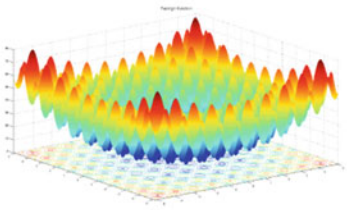
Table 1 Equations and rate of convergence plots for various benchmark functions

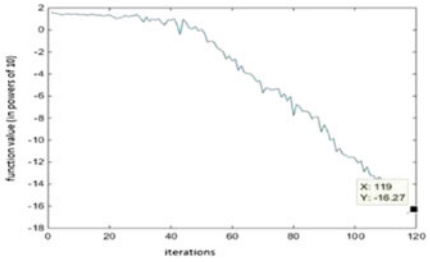
$$SphereFunction : f = \sum_{i=1}^{D} x_i^2 \tag{5}$$



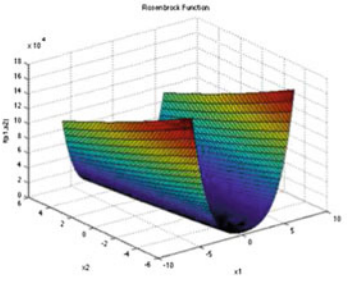


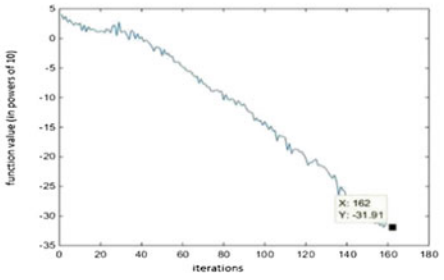
$$RastriginFunction : f = \sum_{i=1}^{D} x_i^2 - 10cos(2\pi i) + 10 \tag{6}$$





$$RosenbrockFunction : f = \sum_{i=1}^{D-1} 100(x_{i+1} - x_i)^2 + (x_i - 1)^2 \tag{7}$$

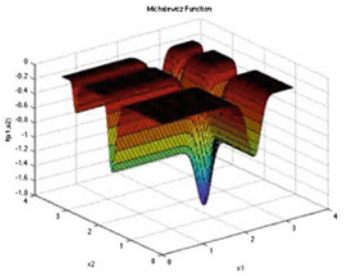


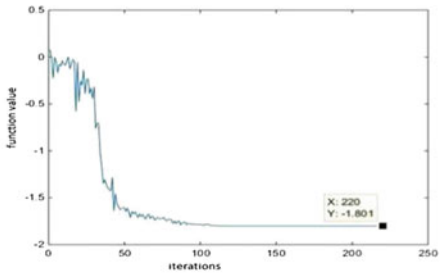


(continued)

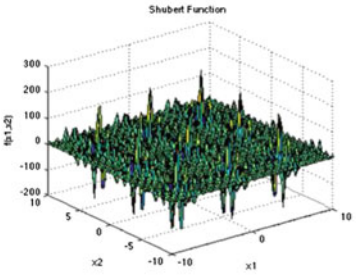
Table 1 (continued)

$$MichalewiczFunction : f = -\sum_{i=1}^{D} sin(x_i) * sin(\frac{ix_i^2}{\pi})^{20} \quad (8)$$





$$ShubertFunction : f = \prod_{i=1}^{D} \sum_{j=1}^{s} jcos((j+1) * x_i + j)) \quad (9)$$



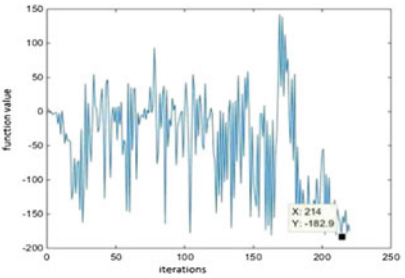


Table 2 Iterations at which benchmark functions converge

Function	Iterations at which the gbest values are obtained			
	Mean	Std. dev.	Fastest	Slowest
Sphere	220	220	220	220
Rastigin	88.0333	6.8956	78	109
Shubert	26.2	2.2652	23	31
Michelawicz	30.0333	3.0904	26	40
Rosenbrock	136.5333	5.3092	128	148

better values in unimodal functions like Sphere function, and almost precise values in multimodal functions like the Rastrigin, Rosenbrock, Michalewicz, etc.

Table 4 contains the comparison of mean and standard deviation on the benchmark functions such as Spherical, Rastrigin, Michaelwicz, and Shubert.

Table 3 Results obtained of the algorithm on various benchmark functions

Benchmark functions		Mean	Std. dev.	Best fit	Worst fit	Expected value [18]
Sphere	X1	–1.45E-37	8.87E-37	6.87E-37	–4.79E-36	0
	X2	–5.60E-37	3.49E-36	2.30E-36	–1.89E-35	0
	Fgbest	1.29E-71	6.54E-71	1.23E-82	3.58E-70	0
Rastrigin	X1	–1.63E-09	2.38E-09	–3.75E-09	3.51E-09	0
	X2	–4.79E-10	2.03E-09	–3.73E-09	2.89E-09	0
	Fgbest	0	0	0	0	0
Shubert	X1	–0.6728	1.5338	–1.4266	4.8581	Several minima
	X2	–0.2959	2.0768	–1.4252	4.8580	Several minima
	Fgbest	–186.725	0.0077	186.7304	–186.705	–186.7309
Michalewicz	X1	2.2029	5.87E-10	2.2029	2.2029	2.20
	X2	1.5708	2.11E-09	1.5708	1.5708	1.57
	Fgbest	–1.8013	6.77E-16	–1.8013	–1.8013	–1.8013
Rosenbrock	X1	1	0	1	1	1
	X2	1	0	1	1	1
	Fgbest	0	0	0	0	0

Table 4 Performance comparison amongst literature and introduced algorithm on benchmark functions

Function name		[11]	[12]	[15]	Obtained results
Mean	Sphere	2.46E-11	1.44E-23	3.80E-27	1.29E-71
	Rastrigin	NA	0.01	0.01	0
	Michalewicz	–1.8769	–1.8947	–1.8966	–1.8013
	Shubert	–186.704	–186.728	–186.717	–186.725
Standard deviation	Sphere	1.35E-10	7.86E-23	2.08E-26	6.54E-71
	Rastrigin	0.1817	0.2524	0.2524	0
	Michalewicz	0.0934	0.0906	0.0868	6.78E-16
	Shubert	0.1418	0.0119	0.0762	0.0077

Rosenbrock function is neglected due to unavailability of information. The best value amongst the compared values is highlighted. On comparison of the mean values with [11] and [15], it is seen that there is 10^{60} increase and a 10^{44} increase, respectively, for sphere function. For Rastrigin function, the expected value of 0 has been obtained. For Michelawicz function, the expected value has been obtained. For Shubert function, [12] exhibits the best value, and the proposed algorithm is only second to it with an error of 0.0016%. In Table 3, it has also been shown that the algorithm produces the expected result for Rosenbrock function.

5 Conclusion

The PSO variant introduced in the paper has three modifications, namely, a new range of linearly varying inertia weight to work with most real-time and natural order functions that follow the exponential rule, a *mutation* technique to control particles that move too fast and increase exploration capability, and a *velocity restriction* factor that converges the search space exponentially over the given range. The algorithm has been proven to work well for both unimodal and multimodal functions. It can especially tackle multimodal functions better due to the inclusion of the Pareto effect in various phases of the algorithm. The algorithm seems to be promising for any number of dimensions with any function and is expected to produce a better solution. Improvement of the order of 10^{40} is seen in spherical function and expected values have been obtained in other benchmark functions.

References

1. Kennedy, J., Eberhart, R.C.: Particle swarm optimization. Proceedings of IEEE International Conference on Neural Networks, Perth, Australia **4**, 1942–1948 (1995)
2. Kennedy, J., Eberhart, R.C., Shi, Y.H.: Swarm Intelligence. Morgan Kaufmann, San Mateo, CA (2001)
3. Eberhart , R.C., Kennedy, J.: A new optimizer using particle swarm theory. In: Proceedings of 6th International Symposium Micromachine Human Science, Nagoya, Japan, pp. 39–43 (1995)
4. Eberhart, R.C., Shi, Y.H.: Particle swarm optimization: developments, applications and resources. In: Proceedings of IEEE Congress on Evolutionary Computation, Seoul, Korea, pp. 81–86 (2001)
5. Ciuprina, G., Ioan, D., Munteanu, I.: Use of intelligent-particle swarm optimization in electromagnetics. IEEE Trans. Magn. **38**(2), 1037–1040 (Mar 2002)
6. Liang, J.J., Qin, A.K., Suganthan, P.N., Baskar, S.: Comprehensive learning particle swarm optimizer for global optimization of multimodal functions. IEEE Trans. Evol. Comput. **10**(3), 281–295 (Jun 2006)
7. Ho, S.-Y., Lin, H.-S., Liauh, W.-H., Ho, S.J.: OPSO: orthogonal particle swarm optimization and its application to task assignment problems. IEEE Trans. Syst. Man Cybern. A Syst. Hum. **38**(2), 288–298, Mar 2008
8. Liu, B., Wang, L., Jin, Y.H.: An effective PSO-based mimetic algorithm for flow shop scheduling. IEEE Trans. Syst. Man Cybern. B Cybern. **37**(1), 18–27 (Feb 2007)

9. Eberhart, R.C., Shi, Y.: Guest editorial special issue particle swarm optimization. IEEE Trans. Evol. Comput. **8**(3), 201–203 (Jun 2004)
10. Zhan, Z.-H., Zhang, J.: Adaptive particle swarm optimization. In: IEEE Trans. Syst. Man Cybern. B Cybern. **39**(6), Dec 2009
11. Shi, Y., Eberhart, R.C.: A modified particle swarm optimizer. In: Proceedings of IEEE World Congress Computation Intelligence, p. 6973 (1998)
12. Chen, T.-Y., Chi, T.-M.: On the improvements of the particle swarm optimization algorithm. Adv. Eng. Softw. **41**, 229–239 (2010)
13. Bansal, J.C., Singh, P.K., Saraswat, M., Verma, A., Jadon, S.S., Abraham, A.: Inertia weight strategies in particle swarm optimization. Proceedings of IEEE International Conference on Neural Network, Perth, Australia **4**, 1942–1948 (1995)
14. Das, S., Abrahamm, A., Konar, A.: Particle swarm optimization and differential evolution algorithms: technical analysis, applications and hybridization perspectives. Stud. Comput. Intell. (SCI) **116**, 1–38 (2008)
15. Anand, B., Aakash, I., Akshay, Varrun, V., Reddy, M.K., Sathyasai, T., Devi, M.N.: Improvisation of particle swarm optimization algorithm. In: International Conference on Signal Processing and Integrated Networks (SPIN). India (2014)
16. Kiremire, A.R.: The application of pareto principle in software engineering. 19th October (2011)
17. Wikipedia. Pareto principle. http://en.wikipedia.org/wiki/paretoprinciple. Accessed March 2016
18. Virtual library of simulation experiments: test functions and datasets. http://www.sfu.ca/~ssurjano/. Accessed March 2016

Multipurpose GPS Guided Autonomous Mobile Robot

Dhruba Ningombam, Abhishek Singh and Kshetrimayum Thoithoi Chanu

Abstract An autonomous robot is an unmanned, self-decision making vehicle that does not require any person controlling it. This document provides the development and implementation of a GPS (Global Positioning System) guided autonomous robot. The robot via GSM module can communicate to the base station for accepting the waypoints and sending parameters of the sensors on-board whenever queried. The GPS module gives the current location continuously to the micro-controller which intelligently determines the optimal path between the current location and the next waypoint until it reaches the final destination. It is employed with various sensors and a robotic arm which can be used for remote surveillance and picking or carrying an object from one place to another.

Keywords GPS (global positioning system) · GSM (global system for mobile communications) · Base station · Waypoints · Navigation · Unmanned NMEA (national marine electronics association)

1 Introduction

The current location provided by the GPS module is set as the initial position and the final position and waypoints are set by the user on a map interface. These waypoints are then sent to the robot via SMS or direct TCP/IP connection (Using

D. Ningombam (✉) · A. Singh · K. T. Chanu
Department of Computer Science and Engineering, Sikkim Manipal Institute of Technology, Rangpo, India
e-mail: dningombam@gmail.com

A. Singh
e-mail: abhishek.singhgtt@gmail.com

K. T. Chanu
e-mail: kshthoichanu@gmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_35

the GSM Module). For each pair of waypoints, the microcontroller then calculates the difference in latitudes and longitudes which is then used to find the required aligning direction, Θ. The robot then aligns itself (differential acceleration technique) and then continues to move forward until either the next waypoint is reached or an obstacle is detected by the ultrasonic sensor. In case it detects an obstacle, the microcontroller decides to move left or right using greedy techniques. The process is repeated until the final position is reached. To Align the robot in correct direction, feedback system is used which monitors current direction and then calculate the difference in current and final requires direction which is used to turn the robot left or right based on an algorithm which decides which decides turning in which direction is more feasible. The robot is continuously in contact to its base station via internet (using the GSM module). It can send the parameters or values of the sensors onboard to the base station in real time which makes it perfect for remote surveillance and monitoring severity of any disaster. Also, it can be reprogrammed remotely using internet to change its path or return back to the initial point from where it started. A Smartphone on-board with 3G video calling and automatic call answering facility makes it possible to see the video in real time with no range limit. The vehicle is also equipped with a robotic arm which makes it possible to pick/drop objects from one place to another giving it a great advantage.

2 Architectural Diagram

An Arduino mega (having ATmega2560 as its microprocessor) is used to control and interface various units. Various sensors are connected to the microcontroller like LM-35 temperature sensor, DHT 11 humidity sensor via its TTL Serial port and SPI bus. Many other sensors can be employed as and when required. A GPS module (Media Tek 3339) is connected via a serial port which helps the robot continuously determine its current location. IC L293D connected to the input/output port of the microcontroller helps control the motor. It communicates to its base station using via SIM-900 GSM modem (Fig. 1).

Software based on the .NET framework is used as the base station to communicate to the vehicle via internet for sending the set of waypoints and to get feedback from the vehicle in real-time including current location of the vehicle and the current parameters of all its employed sensors. The employed arm can be controlled using this application. A map based interface is provided (using Google maps) on which the user can set waypoints by clicking on the desired location. All the actual calculations and the algorithms are implemented on the microcontroller on board. The waypoints can also be sent to the vehicle via SMS through registered mobile number.

The vehicle then navigates through the set of specified waypoints avoiding obstacles on its way until it reaches the final point. Each part of the algorithm is discussed in subsequent sections.

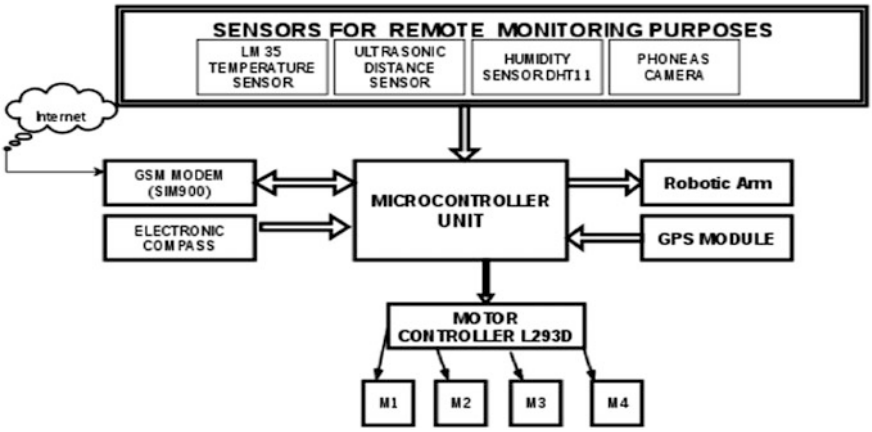


Fig. 1 Block diagram of the GPS based autonomous robot

3 Implementation

3.1 Compass and Direction Aligning

The direction given as output by the compass is given in heading degrees (degree north reference). In this reference, North is considered to be the reference (0°), east is 90°, south is 180° and west is 270° (Fig. 2).

To our conventional Cartesian coordinate system, the degree north reference is related as:

$$\text{cartesian degree} = (450 - \text{compass degree}) \bmod 360 \tag{1}$$

The deciding parameter for vehicle to turn left or right is calculated as:

$$\text{Left} = |(360 - \text{destination}) + \text{current}| \tag{2}$$

$$\text{Right} = |\text{destination} - \text{current}| \tag{3}$$

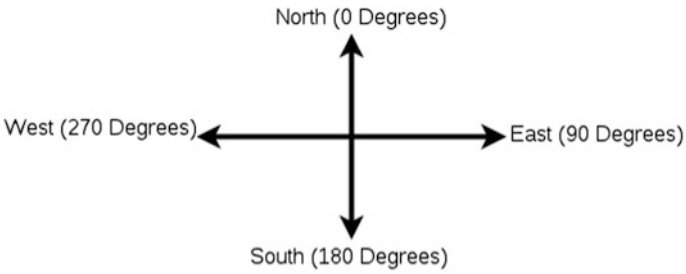


Fig. 2 Relation between heading degrees and cardinal directions

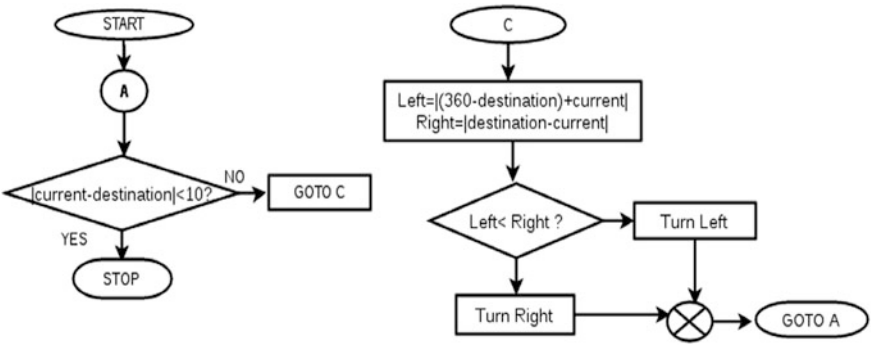


Fig. 3 Flow chart for deciding turning direction

The alignment to destination angle is done as: |current − destination| is the absolute the error between the current heading degrees to the required direction of alignment. Generally the electronic compasses are not 100% accurate, so an error of less than 10° is considered to be fine (Fig. 3).

3.2 *Accepting Way-Points and Finding Path Between Current and Final Point*

We know that latitudes are (imaginary) parallel lines that divide the Earth into grids. The equator is considered as the reference (0°). These lines are equidistant. 1° latitude is approximately equal to 69 miles. 1° is further divided into 60′ (min) and 1′ is further divided into 60″ (s). Hence, 1° latitude = 68.71 mi = 110.57 km, 1′ latitude = 1.15 mi = 1.84 km, and, 1″ latitude = 100.77 ft. = 30.72 m.

Similarly, meridians are imaginary vertical lines further divided in the same fashion as the latitudes. Hence, 1° longitude = 69.17 mi = 111.32 km, 1′ longitude = 1.16 mi = 1.86 km, and, 1″ longitude = 101.29 ft. = 31.03 m

We consider latitudes as X axis and longitudes as Y axis. Current coordinates be (lat1, lon1) and that of next waypoint be (lat2, lon2) (Fig. 4).

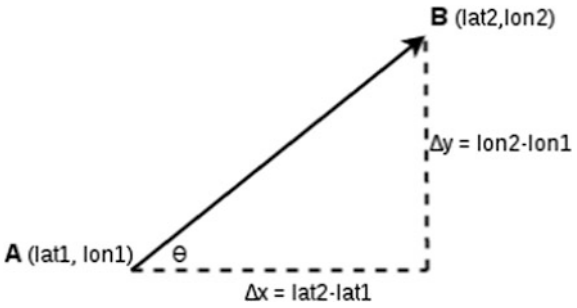


Fig. 4 Calculating angle and shortest distance

We calculate,

$$\Delta x = \text{lat2} - \text{lat1} \tag{4}$$

$$\Delta y = \text{lon2} - \text{lon1} \tag{5}$$

The angle that the robot needs to align is given as:

$$\theta = \arctan(\Delta y / \Delta x) \tag{6}$$

The distance between points A and B is calculated as:

$$\Delta x\text{dist} = \Delta x * 110570 \tag{7}$$

$$\Delta y\text{dist} = \Delta y * 111320 \tag{8}$$

$$\Delta\text{distance} = \sqrt{\Delta X\text{dist}^2 + \Delta Y\text{dist}^2} \tag{9}$$

The number of way-points and coordinates of each waypoint can be initialized by either using the map based software interface or sending SMS in the format specified in the introduction section. The robot then calculates the angle using the formula discussed above and the shortest distance path along the diagonal. It then continues to move forward (Fig. 5).

3.3 Aligning in Correct Direction

After calculating (destination), i.e. $\theta = \arctan(\Delta y/\Delta x)$, if there is error between its current heading angle and Θ, the robot aligns itself in the correct direction (considering error $<10^\circ$ to be correct) using the algorithm discussed in Fig. 6.

3.4 Avoiding Obstacles on the Way

The employed ultrasonic distance meter continuously gives the distance of the nearest obstacle. It has a pair of ultrasonic transmitter and receiver. It works by sending an ultrasonic pulse and finding the time lag between the time it was sent and the receiving time.

$$\text{Distance} = (\text{Speed} \times \text{Time})/2 \tag{10}$$

Obstacle is considered to be present if distance <=10 cm. On detecting an obstacle, it turns the sensor left and right using servo pan over which it is employed and finds whether going left or right is more feasible (Fig. 7).

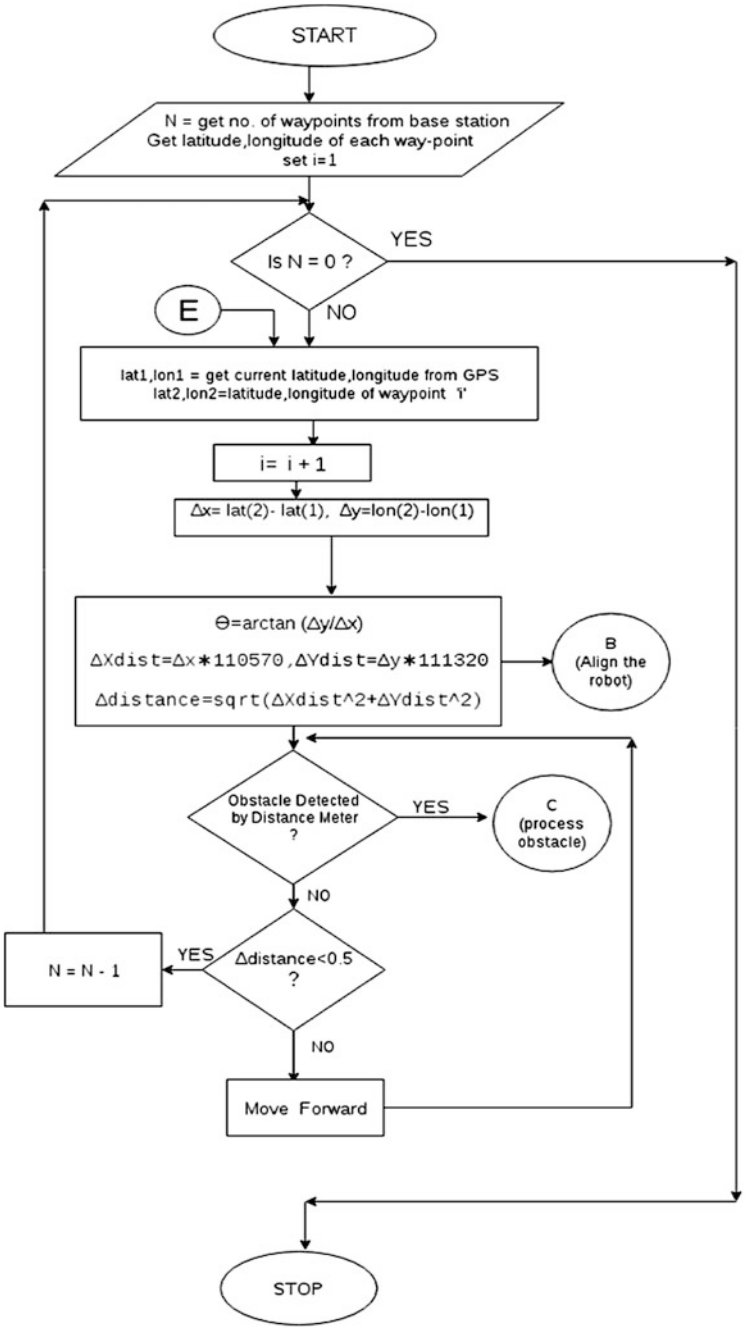


Fig. 5 Decision making and path finding algorithm

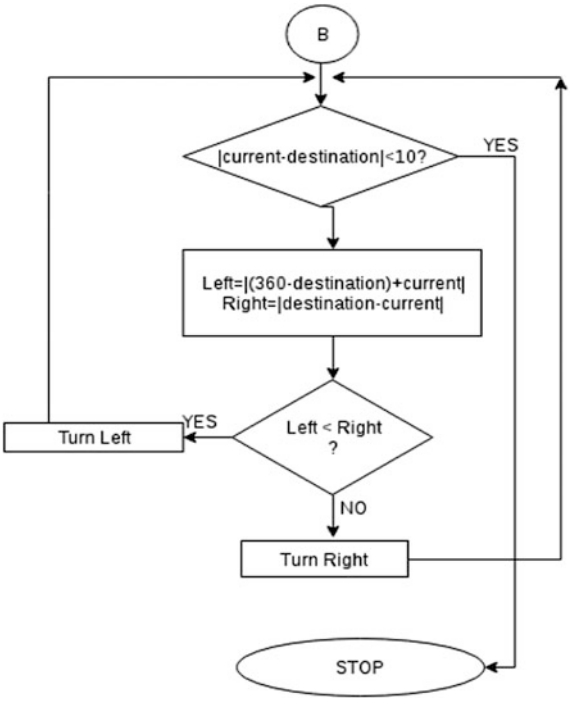


Fig. 6 Flow chart to find optimal turning direction

4 Observations

If point A is the initial position and point F is the final position and the waypoints are B, C, D, E, the robot is observed to follow the following path (Fig. 8).

If any obstacle is encountered between any two way-points, for example between A and B, the robot is observed to align itself as shown in Fig. 9.

5 Applications

- The robot can be used for remote surveillance.
- It can be used to inspect hazardous areas not safe for humans like chemical dump sites.
- It can be used to audit the degree of destruction in places hit by natural disaster.
- It can be used for military purpose by employing weapons.
- It can be used in delivery systems.

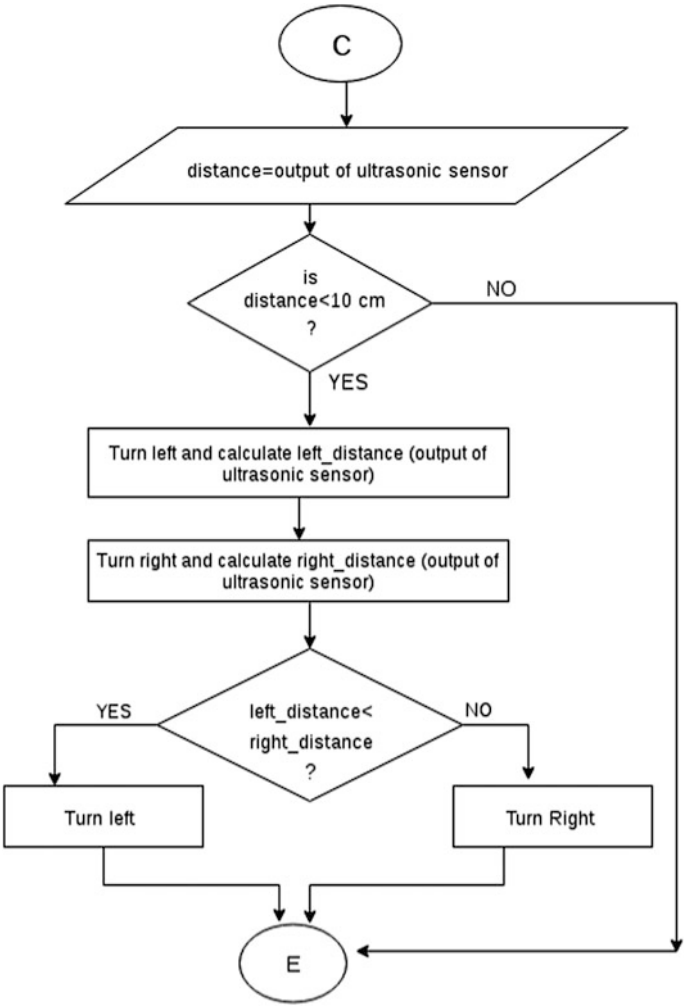


Fig. 7 Flow chart to decide which direction to be turned in case an obstacle is encountered

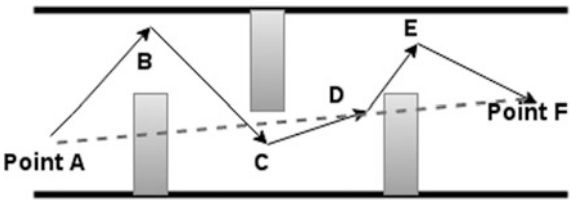


Fig. 8 Dotted line showing straight line path (assuming no obstacles) and the arrows show the actual path followed by robot (blocks representing obstacles)

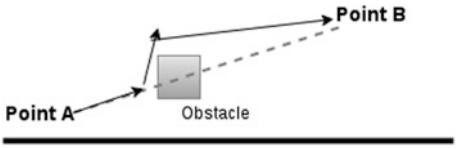


Fig. 9 Behavior of robot under obstacles (arrows indicates path followed)

6 Pseudo Code

6.1 Algorithm: Align (Dangle)

dangle: destination angle to align
current: current heading degrees (output from compass module), updates in real-time

```
1. left = |360-dangle+current|
2. right = |dangle - current|
3. if (!(current >=dangle-5 AND current<=dangle+5)) then
  1. while(!(current >=dangle-5 AND current <=dangle+5)
    1. if(right<left) then
      1. move left
    2. else move right
    3. end if
  2. End while
4. End
```

6.2 Algorithm: Obstacle_Distance

Returns: distance of nearest obstacle in front of Ultrasonic distance meter

```
1. Trigger Transmitter and initialize time, t=0
2. Measure time lag between step 1 and receiver HIGH
(time=t)
3. distance = (330 * t)/2
4. return distance
5. End
```

6.3 Algorithm: Traverse

waypoints [1 … N]: array of waypoints
reached [1 … N]: Boolean array indicating ith waypoint is reached or not
initialize reached[1 … N] as false i = 0

x = current latitude
y = current longitude

```
1. while (reached[N] = FALSE)
  1. dx = x – waypoint[i+1].x
  2. dy = y – waypoint[i+1].y
  3. dist_x = dx * 110570
  4. dist_y = dy* 111320
  5. error = sqrt(dist_x*dist_x + dist_y*dist_y)
  6. do
    1. theta = atan(dy/dx)
    2. align (theta)
    3. obs = obstacle_distance
    4. if(obs<= 10 cm)
      1. avoid_obstacle()
      2. end if
    5. move forward
    6. while(error>1)
    7. set reached[i]=TRUE
    8. i=i+1
    9. end while
2. End
```

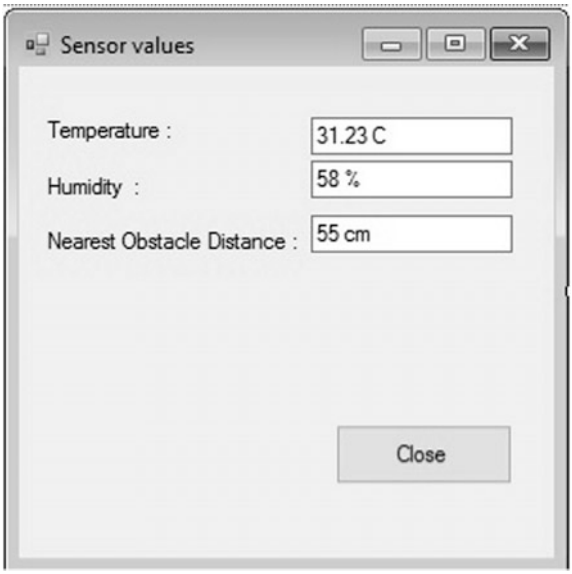
6.4 *Algorithm: Avoid_Obstacle*

Ultrasonic distance meter is fitted on a servo pan/tilt which can move left or right.

```
1. distance = output of ultrasonic sensor
2. if (distance < 10 cm)
  1. turn ultrasonic sensor left
  2. left=output of ultrasonic sensor(left obstacle dis-
tance)
  3. turn ultrasonic sensor right
  4. right =output of ultrasonic sensor(right obstacle
distance)
  5. if(left<right)
    1. turn left
  2. else turn right
  6. end if
  7. goto traverse: step 1.1
3. end
```

Fig. 10 Map based interface for path planning and uploading way-points

Fig. 11 Window showing real time value of the sensors onboard



7 Map Based Software Interface for Setting Waypoints and Real Time Monitoring

The application for base station is written in Microsoft.NET framework. It makes the machine a server (base station) for the robot to communicate over internet using TCP/IP connection. Dynamic DNS is used to provide static hostname over dynamic IP. Free service can be taken from noip.com. The noip client needs to be installed on the server. The application features real-time monitoring of current location and the sensor values as shown in the Figs. 10 and 11.

References

1. Cox, I.J.: Blanche-an experiment in guidance and navigation of an autonomous robot vehicle. IEEE **7**(2)
2. Cox, I.J.: Blanche: position estimation for an autonomous robot vehicle. In: Autonomous Robot Vehicle, pp. 221–228
3. Global Positioning System. http://en.wikipedia.org/wiki/Global_Positioning_System
4. Igoe, T.: Making Thing Talk. O'Reilly Media (2007)
5. Maier, D., Kleiner, A., Improved GPS sensor model for mobile robots in urban terrain. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 4385–4390. IEEE (2010)
6. Malvino, A.P., Bates, D.: Electronic Principles. McGraw Hill (2015)
7. Margolis, M.: Arduino Cookbook. O'Reilly Media (2011)
8. Mester, G.: Intelligent mobile robot motion control in unstructured environments. Acta Polytech. Hung. **7**(4), 153–165 (2010)
9. Olson, M.: Remote control of a semi-autonomous robot vehicle over a time-delayed link. Ph.D. dissertation, University of Saskatchewan (2001)
10. Static name over dynamic IP (dynamic DNS). http://noip.com
11. Thrun, S., Schulte, J., Rosenberg, C.: Interaction with mobile robots in public places. IEEE Intell. Syst. 7–11 (2000)
12. Yuta, S., Suzuk, S., Lida, S.: Implementation of a small size experimental self-contained autonomous robot—sensors, vehicle control and description of sensor based behavior. In: Section 6: Mobile Robots. Lecture Notes in Control and Information Sciences, vol. 190, pp 344–358

A Modification to Graph Based Approach for Extraction Based Automatic Text Summarization

Sunchit Sehgal, Badal Kumar, Maheshwar, Lakshay Rampal and Ankit Chaliya

Abstract The paper lays emphasis on TextRank algorithm, a graph based approach used to tackle the automatic article summarization problem and proposing a variation to the similarity function used to compute scores during sentence extraction. The paper also emphasizes on the role of title of an article (if provided) in extracting an optimal, normalized score for each sentence.

Keywords TextRank · Similarity · PageRank · Lexemes

1 Introduction

Ranking algorithms for undirected graph such as Google's PageRank algorithm [1] have been successfully able to establish their importance and use in social networks and especially the WWW (World Wide Web). Computing the values along the vertices and edges helps one to decide the path which may be an optimal solution to any query.

A similar approach is quite applicable in the field of Natural Language Processing wherein lexical or semantic graphs [2] have been used to extract useful and important phrases from the text available. One such prominent example is the Text

S. Sehgal (✉) · B. Kumar · Maheshwar · L. Rampal · A. Chaliya
CS Department, BVCOE, Paschim Vihar, New Delhi 110063, India
e-mail: sunchitsehgal94@gmail.com

B. Kumar
e-mail: badalkr9@gmail.com

Maheshwar
e-mail: maheshwar1524@gmail.com

L. Rampal
e-mail: kevinpietersanl@gmail.com

A. Chaliya
e-mail: therock110001@gmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_36

Rank Algorithm [2], which is a text oriented ranking based method. The graph based TextRank algorithm is used to extract useful paraphrases and construct a useful and meaningful summary of the text/article available. The algorithm has been a center of research for a long time and has its own limitations too.

A multi document summarization model to reduce the redundancy is discussed in [3]. This model uses the statistical and linguistics for overcoming the information diversity problem. Paper [4] discussed the DBPedia for topic abstraction from clusters of online comments to news. Graph based technique for tweet summarization is used in paper [5]. K-mean clustering algorithm for extraction and text summarization is used in paper [6]. Text categorization for classifying a document in different categories is discussed in [7]. The authors have used KNN based machine learning model for this task. In paper [8] all the text summarization techniques have been discussed.

In this paper, we shall draw our focus towards the TextRank Algorithm and its limitations. Further, we shall present our modifications to the TextRank algorithm and factors (such as the title of an article) that can be incorporated while extracting a meaningful and coherent summary.

The whole paper is divided into five sections. Section 1 being an introduction section and related work done in this direction. Section 2 details the textrank algorithm and how sentence is extracted. Section 3 emphasizes our proposed approach for automatic text summarization. Section 4 shows the implementation part and the results evaluated using our approach. Section 5 summarizes the whole work and direction to future work.

2 TextRank Algorithm

In this section, we will talk about the TextRank algorithm and its salient features. TextRank is an unsupervised machine learning algorithm. It is a type of Extraction based summarization [9], that is, it is used to extract relatively important sentences from each paragraph and arrangement of such sentences to build a relevant summary. The application of TextRank [9] is found in both keyword extraction from a large pool of words and sentence extraction from a body of documents or a single document. It uses a graph based ranking approach wherein each sentence/word represents a node/vertices while the weighted edges represent the degree of similarity of between the vertices. The TextRank algorithm is an extension of the PageRank algorithm where the modified formula is used to calculate the cumulative score of each vertex representing a sentence. However, we propose to modify the similarity function and normalize the scores in order to produce better results. Major advantages of the TextRank [9] algorithm are as follows:

- It is unsupervised, therefore does not require any training set.
- No dependence on language.

The TextRank algorithm is based purely on the frequency of occurrence of words and does not require any prior knowledge of grammar. This eliminates the requirement of any particular tools dedicated to any particular languages. However, this may draw certain limitations to the algorithm particularly in cases of lexemes.

Let G = (V, E) be an undirected graph [2] consisting of set of vertices V and set of edges E ($E \subseteq V \times V$). For a given vertex $V_{i,}$ let In (V_i) represent set of vertices pointing towards the former vertices and Out (V_i) represent the set of vertices that point to the next-inline vertices.

The Score [1] of each vertex is calculated by the formula:

$$\mathrm{PR}(Vi) = (1-d) + d * XVj \in \mathrm{In}(Vi)\mathrm{PR}(Vj)|\mathrm{Out}(Vj) \tag{1}$$

2.1 Sentence Extraction

Applying TextRank consists of building a graph associated to the pool of sentences/ text available where the vertex represents an individual sentence which is to be ranked. All the sentences are ranked in the same way.

For applying TextRank to our text, we first need to build a graph associated [2] with the text, wherein the vertices depict the sentences to be ranked. In order to pick up relevant sentences, we need to rank all of them, and a vertex is created for each sentence in the text.

An edge is added in the graph between two sentences on the basis of the degree of similarity between them which is measured by the degree of words common between the sentences. An edge is connected between a pair of sentences that have common words. In order to avoid promoting long sentences [2], the formula uses a normalizing factor to divide the magnitude of overlapping content between two corresponding sentences. Let there be two sentences, S_i and S_j [2], where a sentence is represented by N_i words that form a sentence:

$S_i = W_{k_1}, W_{k_2}, \ldots\ldots\ldots, W_{k_N}$, the similarity [2] between S_i and S_j is calculated using the below mentioned formula:

$$\mathrm{Similarity}\left(S_i, S_j\right) = \frac{|W_k|W_k \epsilon S_i \& W_k \epsilon S_j|}{\log(|S_i|) + \log\left(|S_j|\right)} \tag{2}$$

3 Proposed Approach

The title may be a name representing the subject in the article, or it may be used to describe a particular situation or description. It is believed that article titles are unique, that is, no two articles can have the same title. Articles titles can also add necessary distinguishing information to elaborate the meaning of the same.

The title of an article, if available, can further help us extract a more meaningful and precise sentence during the process of extraction based summarization of an article. We, therefore, propose to add title as another important factor in the process of article summarization.

While we traditionally used to calculate the similarity between two sentences based on their degree of content overlap between them, the method can be employed while calculating the similarity between each individual sentence and the title of the article as well. The degree of similarity can, therefore be added which shall incorporate the importance of the title for an article as well, and hence, making sentence extraction more meaningful and coherent.

The modified similarity function for comparing two sentences is given by:

$$\text{Similarity}_{\text{sentences}}(S_i, S_j) = \frac{|W_k|W_k \epsilon S_i \& W_k \epsilon S_j|}{(|S_i| + |S_j|)/2} \tag{3}$$

Similarly, the modified function for comparing each individual sentence is given by:

$$\text{Similarity}_{\text{title}}(S_i, S_{title}) = \frac{|W_k|W_k \epsilon S_i \& W_k \epsilon S_{title}|}{(|S_i| + |S_{title}|)/2} \tag{4}$$

Therefore the cumulative score of any sentence, S_i, say, S_1 is given by:

$$\text{Similarity}_{\text{title}}(S_1, S_{\text{title}}) + \{\sum_{i=1, j=2}^{j=N} \text{Similarity}_{\text{sentences}}(S_i, S_j)\} - \text{Similarity}_{\text{sentences}}(S_1, S_1) \tag{5}$$

4 Implementation and Results

Implementation is the developmental stage of the theoretical design [10]. At this stage, the project is turned into a working system. The total implementation of the project is divided into two important modules:

Module 1 Uploading of input file which contains the article, processing on the input text file and calculation of scores

The file is uploaded using a dialog box and the title of the article is specified in the text input. Splitting of article into sentences and two stages of comparison take place:

- Comparison of each sentence with every other sentence.
- Comparison of each sentence with title.

The cumulative score of each sentence is assessed and the graph matrix is created.

Module 2 Choosing best sentences from each paragraph and displaying the summary in the output window

The best sentences are selected from each paragraph and the summary is displayed on the output window.

We applied the modified sentence extraction formula on multiple articles for summarization task and evaluated the results. We took nearly 4 sample articles for article summarization and evaluated our results using ROUGE 2.0 Evaluation Technique. This method is found to be precisely related to human evaluation as it is based on Ngram statistics [11]. The Results are further mentioned in a table given below.

5 Conclusion and Future Scope

The paper introduces the TextRank Algorithm which is an Extractive Summarization technique based on undirected graphs. We also talked about the way in which sentences are extracted and the importance of a title in an article summarization. We, further devised a formula for the same, and illustrated how our implementation really worked. Engaging title in the process of summarization of the article (if available) ensures consistency and coherence and that the best suitable candidate/sentence is extracted in accordance to the sense of the article. The results of four sample articles were computed using ROGUE 2.0 evaluation toolkit based on several parameters, as depicted in Table 1.

While efforts have been made to extract a meaningful and coherent summary from the article, there is still a lot of scope of improvement as to how the sentences are extracted and whether they take the summary to its logical meaning. Considering various other factors like personal pronouns, lexemes [12] can further ensure a meaningful, logical and coherent summary of an article.

Table 1 Results of summary evaluation using ROUGE 2.0 Evaluation Toolkit. The summary is generated using the modified sentence extraction formula

Rouge type	Task name	Average recall	Average precision	Average Fscore	Number referenced summaries
ROUGE 1	Sample 1	1.0	0.29664	0.45732	1
ROUGE 1	Sample 2	1.0	0.09125	0.16841	1
ROUGE 1	Sample 3	1.0	0.33504	0.50192	1
ROUGE 1	Sample 4	1.0	0.44071	0.61180	1

References

1. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine computer networks and ISDN systems, 30(1–7) (1998)
2. Mihalcea, R.: Graph-based ranking algorithms for sentence extraction, applied to text summarization. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Lingusitics (ACL 2004) (companion volume), Barcelona, Spain (2004)
3. Ferreira, R., de Souza Cabral, L., Freitas, F., Lins, R.D., de França Silva, G., Simske, S.J., Favaro, L.: A multi-document summarization system based on statistics and linguistic treatment. Expert Syst. Appl. **41**(13) 5780–5787 (2014). ISSN 0957-4174 https://doi.org/10.1016/j.eswa.2014.03.023
4. Ahmet, A., Emina, K., Balamurali, A.R., Paramita, M., Barker, E., Hepple, M., Gaizauskas R.: A graph-based approach to topic clustering for online comments to news. In: Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016 vol. 20–23, pp. 15-29. Padua, Italy, (2016). isbn-978-3-319-30671-1. https://doi.org/10.1007/978-3-319-30671-1_2
5. Dutta, S., Ghatak, S., Roy, M., Ghosh, S., Das, A.K.: A graph based clustering technique for tweet summarization. In: 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions) pp. 1–6. Noida (2015). https://doi.org/10.1109/ICRITO.2015.7359276
6. Agrawal, A., Gupta, U.: Extraction based approach for text summarization using K-means clustering. Int. J. Sci. Res. Publ. (IJSRP) **4**(11) (2014)
7. Bijalwan, V., Kumar, V., Kumari, P., Pascual, J.: KNN based machine learning approach for text and document mining. Int. J. Database Theory Appl. **7**(1), 61–70 (2014)
8. Mahak, G., Vishal, G.: Recent automatic text summarization techniques: a survey. Artif. Intel. Rev. 1–66 (2016). issn-1573-7462. https://doi.org/10.1007/s10462-016-9475-9
9. Balcerzak, B., Jaworski, W., Wierzbicki, A.: Application of text rank algorithm for credibility assessment. In: Institute of Informatics, University of Warsaw, vol. 2, pp. 02–097. Banacha, Warsaw, Poland
10. Pawar, D.D., Bewoor, M.S., Patil, S.H.: Text rank: a novel concept for extraction based text summarization. Int. J. Comp. Sci. Inf. Technol. (IJCSIT) **5**(3), 3301–3304 (2014)
11. Ganesan, K., Zhai, C., Han, J.: Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 340–348 (2010)
12. Halliday, M., Hasan, R.: Cohesion in english. Longman (1976)

Intellectual Conveyance Structure for Travellers

Vishal B. Pattanashetty, Nalini C. Iyer and H. L. Viswanath

Abstract Intellectual conveyance structure for travellers is gaining credit in emerging nations like India. This paper presents the new structure that can be used to forecast the estimated distinct bus stops and alerts the travellers when their stops arrive as well as any occurrences like fire hooks in the bus. This paper details the system hardware and software design, the functionality of the system. The involuntary real-time traveller alert information structure has the impending of generating the conveyance structure which is a smart alternative for the people who travel long journey and facilitating to less private automobiles on the lane which providing reduced embarrassment while travelling a long journey. This paper deliberates the gains of using ZigBee wireless communication 802.15.4 (WSN) technology as a transit controlling and planning aid. It is shown how ZigBee wireless communication 802.15.4 (WSN) module can be used for efficient gathering and examination of passenger movements, thus providing beneficial information for transit operations management and short-term planning. Moreover, in combination with intellectual conveyance structure technology, ZigBee wireless communication 802.15.4 (WSN) can also deliver reliable, very low cost, self-healable, sustenance for infrastructure and fleet administration activities, as well as, existent time statistics for transit user.

Keywords ZigBee · 802.15.4 · Wireless communication · Intelligent transport system (ITS) · Real-time passenger information system

V. B. Pattanashetty (✉) · N. C. Iyer · H. L. Viswanath
Department of Instrumentation Technology, B.V. Bhoomaraddi College of Engineering and Technology, Vidyanagar, Hubballi, Karnataka, India
e-mail: vishalbps@gmail.com

N. C. Iyer
e-mail: nalini_c@bvb.edu

H. L. Viswanath
e-mail: hl_viswa@hotmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_37

1 Introduction

In the current situation, individuals like to travel a long journey through reserved buses or train, and most of them frequently prefer night journey. At the time of journey people may sleep in the bus or train, in many cases, they cannot be alert of their stops. At each destination, the conductor comes and announces the respective destination thereby disturbing the sleep of the other passengers as well as the environment. How would it be if the specified person is alerted at his respective destination without disturbing the other passengers? In present days there have been many alterations in the field of transport in India may be due to, higher expectations from users, growing economy and maybe in differences of vehicle ownership, most of the public transportation systems do not have the facility to alert passenger after reaching their respective destination. To overcome these problems many systems were implemented. The current study and implementation exertions are habitually oriented towards a traditional scenario.

Reference [1] describes some of the current passenger assist systems in diverse portions of the assistance systems. It also relates various mechanisms of vehicle location technology, location estimate and way of the data broadcast. Reference [2] describes particulars around assistance systems project. Reference [3] describes a project study in the system of logical conveyance structure. Reference [4] describes the retaining tourist data of bus travelling systems which use GPS as the main system. References [5–7] offerings a commercially existing traveller data arrangement that takes structures like information sending via GSM and websites. Wireless sensor communication technologies have a major impact on society, widely used in industry, medical, scientific, navy and military applications. CDMA was widely used for wireless communication, which uses code division multiple access technology and has been industrialized for engineering automation. Automation is widely used for transportation applications. Passenger data Structure for travellers in civic conveyance structure or earmarked vehicles by expending wireless communication technology, i.e. by adopting R-F endpoint, R-F endpoint is more effective, consistent and less costly. This exploration is just not only technique. There are multiple ways this solution can be implemented. This exploration evidently defines the proficiency and the competency of R-F. R-F technology is used for making smart vehicles but this analysis delivers an awareness of expending R-F which is used as a communication tool for inter automobile and automobile to infrastructure communication. R-F communication technology is used to interconnect vehicles and it minimizes the entire price of the structure because R-F modules are low-priced than 3G, Wi-Fi and Wi-Max devices (Table 1).

GPS is repeatedly used by nationals, residences and military as a navigation system. On the earth, a GPS receiver is present, it contains a computer system which triangulates data between at least three satellites. The outcome is provided in the form of geographic positions, i.e. longitude and latitude to the nearest receivers, with an accuracy of 10–80 m. Software applications can then use those coordinates to provide driving or walking instructions. Commonly GPS is involved to locate the

Table 1 Wireless connectivity techniques

Parameter	Bluetooth	ZigBee	Wi-Fi
Data rate	1 Mbit/s	20.40 250 Kbit/s	11 and 54 Mbit/s
Range	10 m	10–100 m	Up to 100 m
Networking topology	Ad hoc small networks	Ad hoc peer to peer, mesh and star	Point to hub
Frequency	2.4 GHz	868 MHz Europe, 900 to 928 America, 2.4 GHz worldwide.	2.4 and 5 GHz
Power consumption	Low	Very low	High
Typical applications	Mobile phones, laptops, PDAs	Industrial control and monitoring, PANs	WLAN, broadband

position of the automobile and the information is restructured in the Internet for retrieving. Data is deposited in the Internet and it can gain access to users, it is not good to share the information of all destinationswith all users, the passengers who require information about specific destination should be provided with that specific data and it should be a cost-effective solution.

2 Design Methodology

We frame the problem by clarifying objectives, identifying constraints, establishing functions and gathering the other information needed to develop an unambiguous statement of a client's wishes, needs and limits that is the customer requirements before we begin conceptual design. During long journey(s), a destination usually arrives at the late/early hours, the time when the passengers are asleep. With this, they miss their stops inviting inconvenience. According to the survey, many people have lost their lives when the bus caught fire. There is a need of a system or device which alerts the passenger as they reach their destinations and in times of calamity (fire, etc.).

2.1 Design Attributes

To alert the passengers the system should have automatic operation, less power consumption, it should be affordable with less or no maintenance, easy installation, it must be durable, it must be electric shock resistant, physical range must be less than 80 m, should be fast and accurate, should be cost effective and with less materials used in its making.

2.2 Preliminary Design

In the preliminary design phase, we identify and preliminarily size/estimate the principal attributes of the chosen design concept or scheme. It is more of a technical work. Under this section, we have identified the inputs and corresponding outputs required for our project. Before implementing our design, it is necessary to pen down all details and test it to see the results and rectify design in case of any errors. By eliminating all the undesirable choices, we come up with the best and feasible design.

3 Functional Analysis

3.1 Working

As the passengers enter in the bus and occupy the seats, one small motor or vibrator will be placed under the chair or in the sensible area, passenger information will be stored in the Arduino. ZigBee transceiver will be placed at the destinations, each destination with different ZigBee transceiver. ZigBee transceiver (transmitter and receiver) work under 24 GHz frequencies. Each transmitter will be sending PAN-Id (Personal Area Network) as the bus enters the range of particular frequency (indicates that the destination has arrived), respective ZigBee transceiver receives

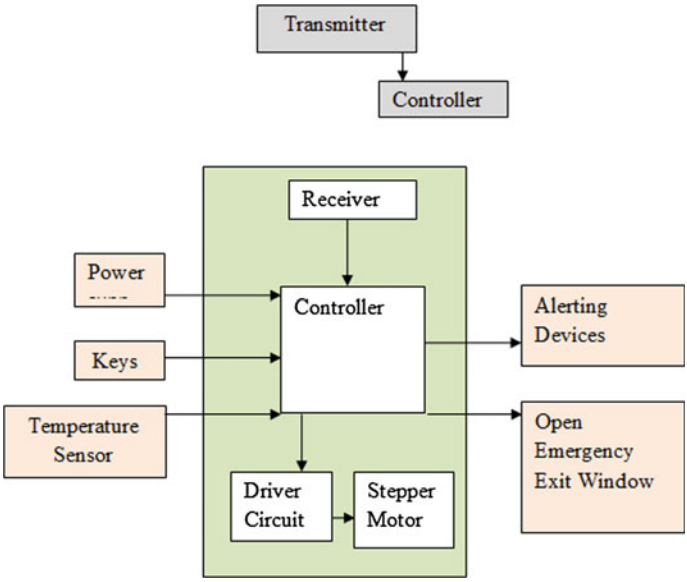


Fig. 1 Functional block diagram

Table 2 Morphological chart

Means Functions	Option 1	Option 2	Option 3	Option 4
Alerting	Low frequency speaker	On/Off of lights	Buzzers	Vibrator with On/Off Of lights (Selected)
Navigation	ZigBee (Xbee) (Selected)	GPS		
Driving Circuit	L293D (Selected)	ULN 2824		
Controller	8051	Arduino (Selected)	Lpc 2148	

PAN-Id or code from the transmitter and hence only the respective seats will be vibrated. During the time of calamity, such as, fire occurrence in the bus or temperature of the bus suddenly rising that time emergency exit window will be opened automatically. LM35 works as a temperature sensor which senses the bus temperature every time, current temperature will be displayed on the LCD (Fig. 1) (Table 2).

4 Results and Conclusion

This section analyzes the results of the developed system. The system development went as expected, resulting with no unusual events that would have introduced constraints, i.e. Passenger assistant system mainly gives alert when particular passenger's destination reaches. Through GPS module only location's latitude and longitude can be tracked. Due to some constraints regarding GPS module, we have switched to ZigBee module which provides accurate and précise result. Thus, considering the above-mentioned problem, we have further proceeded with the work. Design communication includes build models or prototypes to demonstrate or evaluate design's effectiveness (Fig. 4.1).

The proposed system was described, analyzed and compared with the other existing device. The purpose of designing this system is to create a new technique

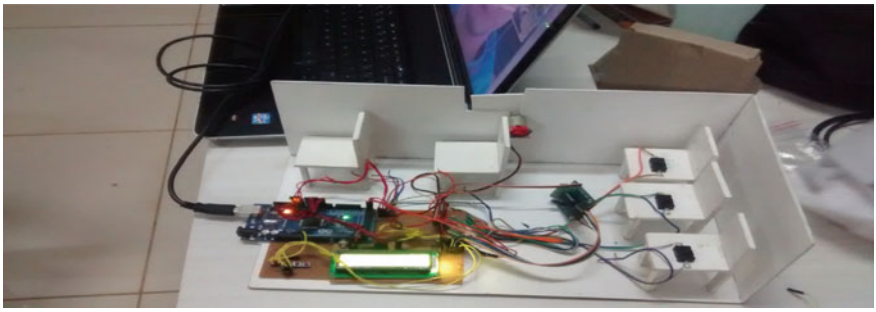


Fig. 2 Design model

which would be superior for real-time operation. It provides convenience to the passengers as well as the drivers. So we conclude to design a system, which works well even under any conditions (Fig. 2).

The future scope of this project is immense, the project is based on ZigBee module which can be further rectified using GPS and RTC, wherein providing the location's information with present day and time which will be helpful for the conductor to feed the present data. Through this project, passenger can only get the alertness when the destination is reached and also the fire alarm.

References

1. Schwinger, C.L.: Real-Time Bus Arrival Information Systems—A Synthesis of Transit Practice. Transportation Research Board (2003)
2. Review of Current Passenger Information Systems, Prepared for the INFOPOLIS 2 Projects (No. TR 4016), Deliverable 1, WP03, Info polis 2 Consortium, August 1998
3. Hu, K., Wong, C.K.: Deploying real-time bus arrival information and transit management systems in Los Angeles. Abstract prepared for the ITS America 12th Annual Meeting, Long Beach, California, April 29–May 2 2002
4. Helsinki Transport System. http://www.hel2.fi/ksv/entire/repPassengerInformation.html
5. Telargo Inc.—Passenger Information Services. http://www.telargo.com/solutions/passener_informationservices.aspx
6. Pattanashetty, V.B., Iyer, N: Smart driving assistance using ZigBee. No. 2015-28-0105. SAE Technical Paper (2015)
7. Kidwell, B: Predicting transit vehicle arrival times. GeoGraphics Laboratory, Bridgewater State College, August 2001

A Viewpoint on Different Data Deduplication Systems and Allied Issues

Shamsher Singh and Ravinder Singh

Abstract Data acts like the heart of an organization, so it needs to be protected from loss and damage. For this purpose, we use backup and recovery strategies. But duplication, which is present in the data, creates a high-cost problem in relation to the storage. With the high rate of increment in the amount of data, it has become a problem to store it in such an efficient manner so that we can reduce the cost of storage and get enough space at low cost. To overcome this problem, at the time of data backup a technique is used to eliminate the redundant data so that one single unique copy of data will be stored which will save space and cost, that technique is known as data deduplication. This paper includes the study of different methods and techniques proposed by other researchers about the data deduplication system.

Keywords Backup · Chunks · Data deduplication · Server · Storage

1 Introduction

In today's era of digital world, storage becomes a very expensive need for the software companies as well as the home users. Everyday millions of users create PB's (Peta Bytes) of data in the form of videos, photographs and other documents. Every user wants its data to be safe from every aspect that is why they store multiple copies of their data at different locations. For this purpose, online storage is provided by many vendors. Another reason is incremental backups that are used for security and consistency means. To overcome this duplication problem data deduplication an effective technique used to free up the storage space.

S. Singh (✉) · R. Singh
Department of Computer Science Engineering, Lovely Professional University, Phagwara, Punjab, India
e-mail: mr.kharal@yahoo.com

R. Singh
e-mail: ravinder.17750@lpu.co.in

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_38

In deduplication, a cryptographic hash is used to locate and delete the redundant data from the backup taken. The hash value is a fixed length output of any data. In deduplication when any data comes for storage, its hash signature is created using a secure hash algorithm (SHA). That hash signature is verified by the server in the hash index which has all the hash signatures stored. If that hash signature matches with any other hash signature it means that data is duplicated and need not store again, then data will be deleted but a reference will be generated to the stored original data. If the hash signature does not match with any other signature, then that data will be stored in the disk and new signature entry will be done in the index.

2 Two Ways to Perform Data Deduplication on Backup

Inline Deduplication: In this data undergoes the deduplication process before storing in the storage disk. When data comes for the storage in the disk the deduplication algorithm is applied on it and only the unique data blocks are stored.

Post Process Deduplication: In this, deduplication is performed on data after storing it into the storage disk. After storing data, data fetched for deduplication process and after that only unique data blocks stored back into the memory and redundant data blocks are deleted [1].

3 Data Deduplication Could Be Done on Three Main Levels

File Level: Which identify the files with different names but having same data in it.

Block Level: Which divides data stream into blocks, and then take its hash signature and matches it with already stored data. For hash signatures, SHA-1 could be used.

Byte Level: It can be called as micro-level deduplication because in this data is divided into bytes and then it goes under the deduplication process [2].

4 Process of Data Deduplication

1. Figure 1a shows the chunking process of the file and deduplication checking process of each chunk. After that, unique data chunks are stored in the disk and duplicated chunks are discarded. Metadata of each chunk is also stored for recovery purpose.

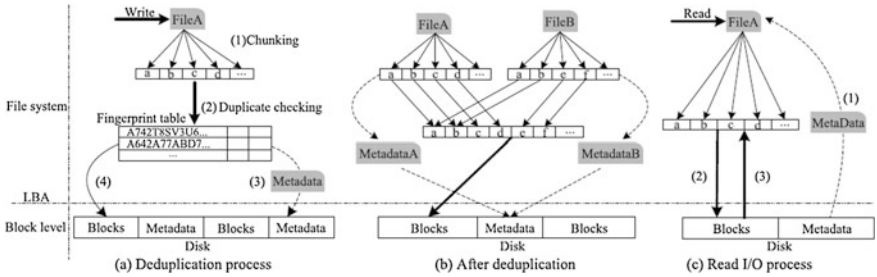


Fig. 1 I/O request process in deduplication system [3]

2. Figure 1b shows that after deduplication process if two files are having some duplicated data then only a single copy of data is going to be stored in the disk and references to that data is made for accessing reasons by File B.
3. Figure 1c shows the read operation of File A. When read request received then by using metadata file is being given to the client who made that request [3].

5 Issues Related to Data Deduplication

Data deduplication is an efficient method for single instance storage. It locates the duplicated data and removes it from storage which makes more free space in the storage system. It ensures that only a single and unique copy of data should be stored in memory. But there are also some issues encountered in data deduplication process. Let us take a tour of them.

5.1 Fingerprint Indexing Issue

In every deduplication system index tables, fingerprint tables and mapping tables are essential. An effective index table design becomes the reason of greater performance of the system.

In traditional deduplication systems index tables consumes very high rate of memory for storing the fingerprints of data blocks just because they used to store the full hash value of every data block.

Multi-level indices with three key tables can be used to resolve the above-said issue: LBA Remapping Table (LRT), Multi-Index Table (MT) and Hash node Table (HT).

Figure 2 illustrates the structure of multi-level indices. LRT store LBA (Logical Block Address) of each entry and pointer to the information of multi-index table. The multi-index table is divided into many cells called buckets. Rather than storing

full fingerprint value buckets are used to store first three 8 bits of fingerprint value. Each bucket stores 256 values with the similar hash prefix. At the last level 136 bit's hash value is stored in the hash table.

In this way by storing fingerprints in multi-level storage, storage space is saved up to some extent for actual data storage.

This index structure is used in a better data deduplication system called I-sieve. It is useful to uplift the performance and reduce the consumption of RAM by using multi-level cache and index and mapping tables.

Figure 3 shows the architecture of I-sieve, in this all the clients interact with storage services using iSCSI protocol. I-sieve acts as a sieve and eliminates the redundant data that is present in the file system. I-sieve consists of three main components: Deduplication Engine, Multi-Cache and Snapshot Module.

Data deduplication engine gives services for block-level deduplication and management of fingerprint table. All the write requests are being handled by duplication engine and afterward forwarded to the disk.

Multi-cache manages all the cache-related tasks and acts as the connector between the memory, solid state drives and storage disks.

Snapshot module periodically takes a snapshot of data blocks for reliability reasons [3].

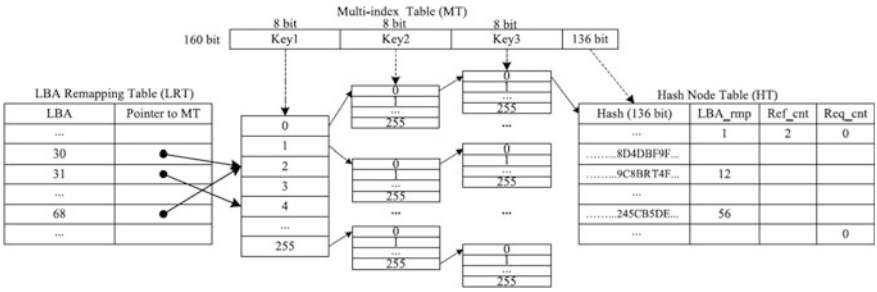


Fig. 2 Index table structure and mapping process [3]

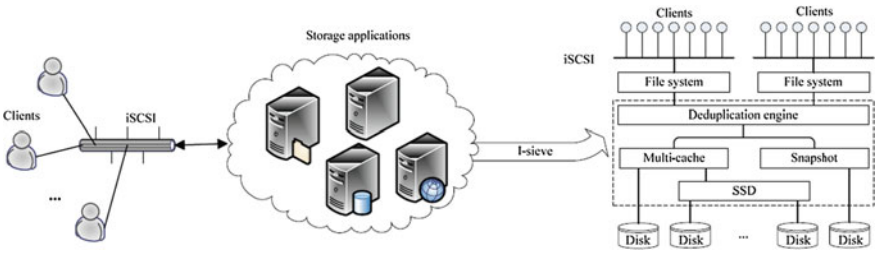


Fig. 3 Architectural overview of I-sieve [3]

5.2 Cross File Similarity Checks

Widely used backup type is file-based backup because in the original world all data is normally stored in the form of files. But file-based backup has some defects, like, it does not make check for cross-file likenesses. As the result, every backup contains similar data which reduces the storage space and increases the requirement for network bandwidth. Cross-file likeness means if in storage disk two files are stored with different names but with identical content. If one makes some little changes in one copy of that data, then every time both the copies will be backed-up. Since granularity of file varies from many bytes to many gigabytes.

A new chunking technique name Anchor-Based Chunking is used to remove the above-said issue. Firstly, this chunking technique is used in FBBM (File-Based Backup Method).

Anchor-based chunking: FBBM divides the data into variable-sized parts according to the content of files. At the point when the low-order n bits of a section's hash value equivalents to a pre-decided value (example 61), the portion constitutes an anchor. A single file may hold numerous non-overlapping anchors. These stays are utilized as chunk boundaries to partition the document into variable-sized lumps. We state to this technique as the anchor-based chunking system. The normal chunk size is controlled by the parameter n which is the quantity of bits used to detect an anchor. For FBBM, n equivalents to 13, so the probable chunk size is 2^{13} B (equals to 8 kB).

Figure 4 demonstrates the structural design of FBBM. This contains three parts: Storage Server (SS), Backup Agent (BA) and Catalog Database (CD). Backup agent could be any program which is installed on the system and to be backed-up. When BA have to backup any file then it divides the documents into variable chunks with the help of anchor-based chunking method and computer the fingerprints of each chunk, then it sends all the fingerprints to SS which respond to it by asking about the fingerprints which are not found in it. After that SS is the responsible part for storing that chunks plus their indices. At the time of restoring process SS will rebuild the file according to the indices stored. Catalog database is used to saves the metadata of the backup for the managing reasons [4].

5.3 Low Performance While Restoring the Data

Data is stored on storage disk in a scattered manner and physical location of data blocks are not in a sequence. It becomes a performance issue when any user demands for the restoration of the data backup taken earlier.

Backup restoration is an important operation in data base systems. It should be completed in an efficient manner to improve the performance of the system. Figure 5 shows the file level data deduplication process. In this after taking data files hash signatures of files are calculated and then unique hash values are stored in

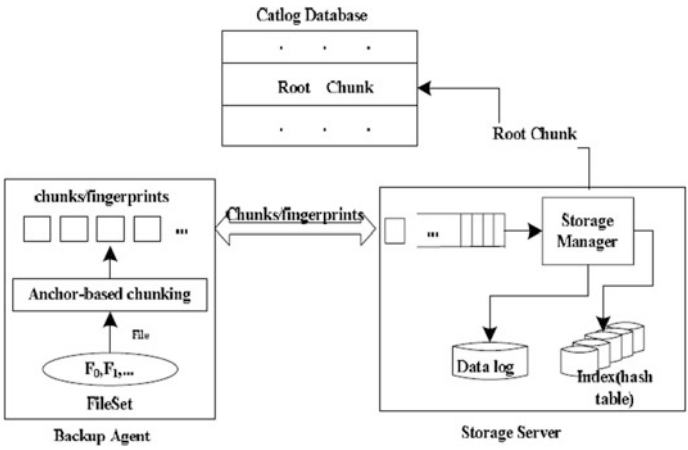


Fig. 4 Architecture of FBBM [4]

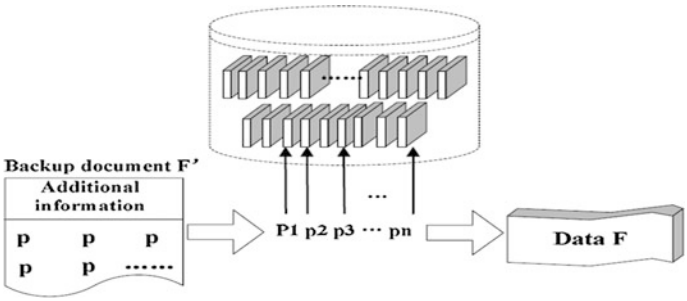


Fig. 5 Recovery based on backup document [2]

index with pointers to the physical location of the file in the storage disk. Files are then stored in physical storage and a **backup document** file is created at that time which will be helpful while recovery.

Recovery based on backup document: When client requests for recovery of data then backup document file is used which contains pointers to the physical location of the files. Using that file all the requested data will be collected and handed over to the client [2].

5.4 Data Fragmentation Issue During Restoration of Backup

In database backup systems data blocks are physically spread after applying deduplication process and creates fragmentation issue.

Chunks that are scattered in memory are the causes of fragmentation. This fragmentation causes of the decrease in restore performance and invalid chunks if user deletes expired backups because at the time of restore chunks should in collected again in the container according to the recipe generated at the time of backup.

To shrink this problem History Aware Rewriting Algorithm (HAR) and Cache Aware Filter (CAF) is proposed. HAR use historical facts to locate and reduce the sparse container. CAF use restores cache information to recognize out-of-order container which causes of restore performance.

During backup data is split into small fixed or variable length chunks and then SHA-1 is used to find out their fingerprints. During backup process, these chunks are aggregated into containers and a recipe of the sequence of these chunks is generated which will be used for restoring the data when needed.

Sparse Container: Like presented in Fig. 6, second backup referenced only single chunk in container IV. So pre-fetching of the fourth container is inefficient for chunk J when restoring the backup number second. After deletion of the first backup, it requires a merging process to recover the invalid chunks in the fourth container.

Out-of-order Container: When any container is call up a number of times irregularly during the process of restore, then we call it as an out-of-order container for the process of restore. As presented in Fig. 6, the fifth container will be called up three times irregularly while the restoring process of the second backup.

For correctly recognize and decrease sparse containers, it observed that the sparse containers stay sparse in the next backup, and therefore suggest HAR. It suggestively increases the performance of restore process with a minor drop of deduplication ratio.

Cache Aware Filter (CAF) proposed to adventure cache data to recognize the out-of-order containers that will destroy the performance of restore. We use it in the hybrid scheme to enhance the performance of the restore process under limited restore cache without a major reduction of deduplication ratio [5].

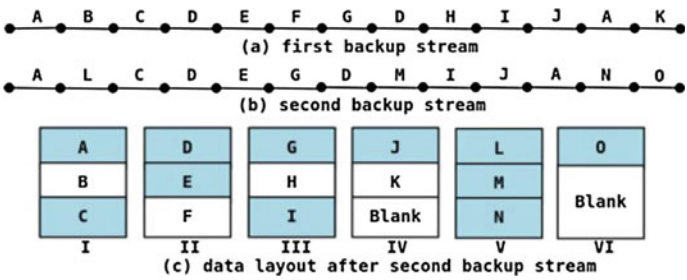


Fig. 6 An example of two successive backups. The colored areas in every container denote the chunks needed by the second backup [5]

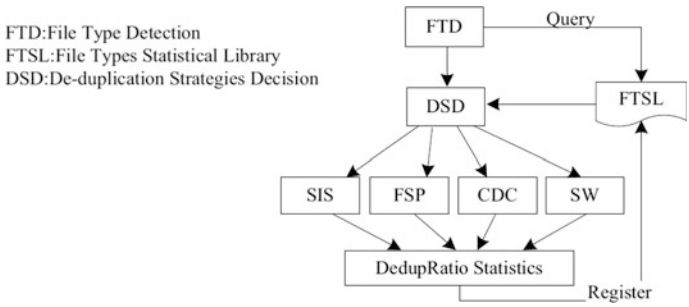


Fig. 7 Intelligent data deduplication strategy [6]

5.5 Performance Issues During Deduplication Process

With the high rate of increment in the amount of data, performance has become a problem to store it in such an efficient manner so that we can reduce the cost of storage and get enough space at low cost.

An intelligent data deduplication-based backup system could be used to eliminate performance issues, in which four different deduplication strategies are used, i.e. SIS, FSP, CDC and SW. These strategies are used for source-side deduplication and these are being used according to the type of data. It provides high trust and security in process of data deduplication.

Intelligent deduplication chooses the appropriate strategy according to the file type and application context. SIS is used for the email attachments and file systems. FSP, CDC and SW are for office applications. It decreases the dependence upon the CPU, increases the deduplication speed and lowering the network bandwidth consumption.

Figure 7 depicts the working of intelligent data deduplication. When any file comes for storage firstly FTD (File Type Detection) detects the type of file and then generate a query for FTSL (File Types Statistical Library) to check whether the file type exists or not. FTSL respond back and after that DSD (Deduplication Strategies Decision) selects the deduplication strategy according to the deduplication ratio status of all the strategies [6].

6 Conclusion

Deduplication of data becomes a key research area of the industry because of the explosive increase in the need of storage space. It eliminates the repeated copies or chunks or segments of data from the storage and replaces it with its pointer. In this digital world data deduplication providing a good way to eliminate the repeated storage of same data and give a huge free space to the users. In every deduplication

system index tables, fingerprint tables and mapping tables are essential. An effective index table design becomes the reason of greater performance of the system. It ensures that only a single and unique copy of data should be stored in memory.

References

1. Vikraman, R., Abirami, S.: A study on various data de-duplication systems. Int. J. Comput. Appl. **94**(4), 35–40 (2014)
2. Sun, G.-Z. et al.: Data backup and recovery based on data de-duplication. In: International Conference On Artificial Intelligence And Computational Intelligence, pp. 379–382 (2010)
3. Wang, J., et al.: I-sieve: an inline high performance deduplication system used in cloud storage. Tsinghua Sci. Technol. **20**(1), 17–27 (2015)
4. Yang, T., et al.: FBBM: a new backup method with data de-duplication capability. In: International Conference on Multimedia and Ubiquitous Engineering, pp. 30–35 (2008)
5. Fu, M., et al.: Reducing fragmentation for in-line deduplication backup storage via exploiting backup history and cache knowledge. IEEE Trans. Parallel Distrib. Syst. **27**(3), 855–868 (2016)
6. Zhu, G., et al.: An intelligent data de-duplication based backup system. In: 15th International Conference on Network-Based Information Systems, pp. 771–776 (2012)

Improved Genetic Algorithm for Selecting Significant Genes in Cancer Diagnosis

Soumen Kumar Pati, Saptarshi Sengupta and Asit K. Das

Abstract Microarray technology serves as a very helpful tool in measuring expression levels of genes the numbers of which range in the thousands. The most challenging issue in classification of cancer using microarray data sets is in selecting the least number of significant genes capable of maximizing accuracy of classifier. Here, we have presented an improved genetic algorithm to select significant genes for cancer diagnosis. A new mutation technique called proximity mutation is used in the paper to preserve the diversity in the population of the genetic algorithm. The term proximity mutation is proposed as distance plays a role in determining the chances of mutation. The fitness function is defined here based on both minimum number of genes and maximum accuracy measured by SVM classifier. Finally, the proposed technique is applied on well-known and publicly available microarray data sets in order to establish the effectiveness of the proposed methodology and provide comparative analysis such that our approach can be examined in light of existing methods.

Keywords Microarray dataset · Gene identification · Genetic algorithm Proximity mutation · SVM classifier

S. K. Pati (✉) · S. Sengupta
Department of Computer Science/IT, St. Thomas' College of Engineering & Technology, 4 D. H. Road, Kolkata, India
e-mail: soumenkrpati@gmail.com

S. Sengupta
e-mail: ssengupta8@hotmail.com

A. K. Das
Department of Computer Science and Technology, Indian Institute of Engineering Science and Technology, Shibpur, Howrah, India
e-mail: akdas@cs.iiests.ac.in

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_39

1 Introduction

In recent times, several applications across myriad fields have seen the production of large quantities of high dimension objects [1] in the wake of various investigations. A recurring observation when dealing with any microarray dataset is that they possess a large number of genes for several hundred or more of the samples present in the set. Using microarray technology [2] we are able to view an enormous number of genes present in the given dataset and derive important insights about the data present in it. This cancer categorization experiments are accepted to give an enhanced catalogue of cancer where a logical analysis of the correlation of expression pattern [3] of high volume of genes to definite phenotypic modification. In DNA microarray technology [2], gene expression levels from tissue samples, obtained from patients, are computed by biologists who investigate and infer how the genes of one patient related to the type of cancer they were diagnosed with. Almost all of the genes examined were strongly related to the cancer type with which the patient was afflicted. But biologists are interested in zooming in on i.e. selecting a small portion of genes which are foremost responsible for the cancer before proceeding to perform further analysis and costly experimentation. Genetic Algorithms (GA) [4, 5] have been used in a great variety of applications as well as foremost solving optimization problems [6]. The results of such problems can never really be said to be correct but can be said to be optimal or best fitting. While studying optimization problems [6] using GA, it is necessary to hone in on the global optimum of the dataset as fast as possible without stagnating at any one particular point. The paper [7] provided a novel method to avoid premature convergence called selective mutation. The paper [8] demonstrated great accuracy in classification with their algorithm using only four genes. In the paper [9], a novel approach that merge the gene ranking method and clustering algorithm selects biomarker genes with fairly acceptable results. The work described in [10] is a rough set theory related soft computing method, where single or even double genes are obtained for cancer classification. In [11], an improved GA based gene selection and SVM classification is done which gives better accuracy but at the expense of more than 15 genes per trial on an average. A fuzzy rule-based gene selection method is proposed in [12] which achieved better accuracy with more number of genes. In [13], a multiple-filter-multiple-wrapper (MFMW) method is reported to identify potential biomarker genes by using more than one filters and wrappers to improve the accuracy.

When gene selection is done using GA, we often deal with large amounts of genes from microarray datasets. The genes in these sets are encoded as binary string in the form of chromosomes which in turn are put through the algorithm. For this purposes of sample classification, we wish to know which samples are cancerous and which are non-cancerous. Under the proposed methodology, the classifiers employed were

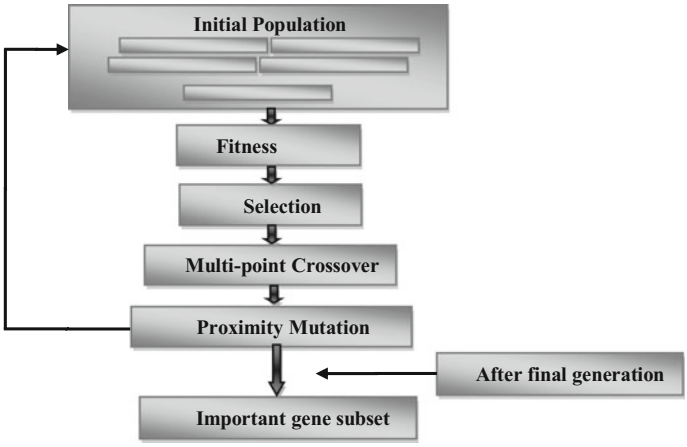


Fig. 1 The proposed gene selection methodology

able to achieve high levels of classification accuracy even in spite of the fact that using our method rendered the quantity of active (selected) genes in a chromosome to an extremely low number. In this paper the proximity mutation was inspired by Selective Mutation but aims to aid in gene selection techniques. The method builds on the idea of single flip bit mutation but greatly enhances it. Here, the distance between two inactive genes (encoded as zeroes) are determined and if it is found that it is less than a certain predefined value, then only they are converted or "flipped" to active genes (ones). The best chromosome was determined from the chromosome pool and a reduced microarray dataset was obtained from it. Finally an SVM classifier was run on the reduced dataset and a very high accuracy in cancer classification was obtained. The proposed gene selection method is highlighted in Fig. 1.

Rest of the paper is organized as follows. Section 2 explains the gene selection methodology based on genetic algorithm. Section 3 discusses the performance of the proposed method and details several experimental results for different gene expression datasets. The paper is ultimately concluded in Sect. 4.

2 Gene Selection Using Genetic Algorithm

In this section the gene selection methodology using genetic algorithm [4, 5] for minimum number of gene selection with maximum classification accuracy of the microarray dataset is proffered. The proposed method selects global-best chromosome from the population using a suitable fitness function computed by the minimum number of genes providing maximum classification accuracy.

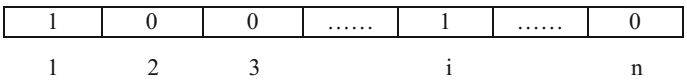


Fig. 2 A randomly generated chromosome

2.1 Initial Population

The initial population are generated randomly consisting of *N* chromosomes. These chromosomes are essentially strings of binary zeros and ones. Each bit (allele) is representative of a gene and each chromosome is of the same length or dimension equal to that of the number of genes in the dataset. A '*1*' in the *i*th position of the chromosome indicates that the *i*th gene from the dataset is active in the considered chromosome while a '*0*' indicates the absence of that particular gene. The entire population is also called the chromosome pool (Fig. 2).

2.2 Parent Selection

The *n*-th number chromosome ($1 \leq n \leq N$) is selected as the first parent and the other parent is randomly selected from rest of chromosomes. In the subsequent iterations, one of the parents for the next generation is the local best chromosome generated in the previous generation while the other was randomly selected from the population.

2.3 Fitness Function

Our objective is to select the best chromosome from the GA population that gives maximum classification accuracy and minimum numbers of genes based on defined fitness function. The fitness function uses two different measures: (a) Classification Accuracy and (b) Minimum Number of Genes. The microarray dataset has two classes of specimen, one being cancerous and the other non-cancerous. The fitness function is defined on the classification accuracy of SVM classifier applied on data subset corresponding to the chromosomes. The SVM classifier is a function based classifier and more effective in two class system and thus makes sense for using it in our work. Our target is to find the minimum number of significant genes capable of generating maximum classification accuracy and as such, the fitness function is

modelled as a linear combination of classification accuracy (*CA*) and number of genes in a chromosome (*NC*). As *CA* value may dominates the *NC* value, so *CA* and *NC* values are normalized into (0, 1) and finally, as accuracy is our main concerned so a weight factor α is assigned to *CA* and β is assigned to *NC*, where, $\alpha > \beta$ and $\alpha + \beta = 1$. Thus, the fitness function is defined in Eq. (1).

$$FV = \left(\alpha \times \frac{CA}{100}\right) + \left(\beta \times \left(\frac{S - NC}{S}\right)\right) \tag{1}$$

where, *S* is the cardinality of genes of the dataset. If *NC* is small and *CA* is high for any chromosome, then *FV* value is high and gives better solution with respect to fitness function. So, maximum of *FV* value allows us to select a better chromosome with respect to maximum accuracy and minimum number of genes.

2.4 Reproduction

Reproduction serves to introduce genetic diversity in the existing chromosomes. In the absence of reproduction, the GA would stall at a very early point or generation and would lead to problems such as premature convergence. Furthermore, without this stage of the GA, the exploration and/or exploitation of the search space would be drastically reduced. Hence reproduction is of the essence in the algorithm. It is carried out via the 2 genetic operators i.e. crossover and mutation.

2.4.1 Multipoint Crossover

As mentioned earlier, the crossover operation is a part of the reproduction phase of the GA responsible for exploration of the solution space. In order to generate fresh new individuals (chromosomes), the crossover function is employed. The outcome of this phase is in the production of two new chromosomes (children), from a pair of chromosomes selected as parents, generated by performing the crossover operation with a probability c_p. In the proposed approach, we have employed the two-point crossover technique. Two-point crossover relies on generating two random points each of which indicates a particular bit position for either parent. The portion of the parent chromosomes lying between the two points are then exchanged thus giving rise to two children chromosomes. The intuition backing two-point crossover is that the children produced after the phase resemble their parents to a greater degree than had we been using the simple one-point crossover.

As such, it becomes obvious that the chromosome pool becomes filled with high quality individuals and thus we can expect convergence to happen at an earlier time.

2.4.2 Proximity Mutation

In single-bit mutation, a gene is randomly selected to be mutated and its value is changed (mutated) depending on the encoding type used. But it suffers from a drawback in that it fails in generating a diverse population as the first bit of the chromosome usually avoids getting altered (mutated). In multi-bit mutation, multiple genes are randomly selected for mutation and there values are changed depending on the encoding type used. So, both of the mutation is depended on the flip-bit mutation and random bit number generation with respect to mutation probability m_p, which is inefficient in high dimensional space. Finding the minimum number of active genes in the chromosome is one of the objectives of our proposed method, so flip-bit mutation methods may diversify the population. To overcome these demerits, proximity mutation is used in the paper for mutating the chromosomes, which builds on flip-bit mutation but modifies it greatly to produce fittest offspring. The proposed mutation method works in the following manner.

Let, two random positions are generated and count the number of '0's and '1's between these positions, say C_0 and C_1, respectively. Then we have two cases to analyse:

Case 1: If $C_0 \leq C_1$, then '0' and '1' are swapped between the selected position and a new offspring is generated.
Case 2: If $C_0 > C_1$ and $(C_0 - C_1) \leq C$, a predefined threshold, then they are flipped, else the chromosome remains unchanged.

These two cases ensure the reduction in number of active genes (i.e., '1's) in the chromosome, which would generate a greater diversity in terms of lower number of '1's compare to flip-bit mutation, which is one of the objective in our work.

2.5 Proposed Algorithm

The proposed gene selection methodology not only gives the maximum classification accuracy but also produces minimum number of genes. The algorithm is defined below:

Procedure: **Gene Selection**

Input: Experimental microarray dataset

Output: The Global-best chromosome with significant genes

Begin

Randomly generate initial population P of size *N;*

Evaluate fitness value of all chromosomes;

Select the global-best chromosome;

Repeat

For (i=1 to *N*) **Do**

First_ parent (P_i)= ith chromosome in the P

Select other parent (P_j) randomly from P;

Apply two-point crossover with probability c_p and create two offspring;

Apply proximity mutation to the offspring with probability m_p;

Evaluate fitness value of the offspring;

If (Fitness value of any offspring is above the global-best) **then**

offspring becomes global-best;

Else if (Fitness value of the offspring dominates its parent(s)) **then**

It replaces the parent of poor fitness value by the offspring;

Else Discard the offspring from the population;

End-For

Until (Predefined number of generations are exhausted)

Return Genes with the global best chromosome;

End.

3 Experimental Results and Performance Evaluation

3.1 Dataset Description and Parameters

We carry out our experiments on benchmark microarray datasets collected which are widely available [14] and contain high volumes of unwanted genes with random noise and the samples are linearly inseparable. Statistics of the microarray datasets are shown in Table 1.

The parameters which are used in the proposed work are outlined in Table 2. After several runs of the proposed algorithm on different datasets, these parameters were finalized upon.

3.2 Performance Analysis

We run the algorithm several times and after final generation minimum (i.e., *Min.*), maximum (i.e., *Max.*), mean (i.e., *Avg.*) and standard deviation (i.e., *Std.*) among all chromosomes in the population are measured in terms of fitness value, listed in Table 3.

Table 1 Statistics of the microarray dataset

Dataset	Gene no.	Class name	Sample no. (class1/class2)
Leukaemia	7129	ALL/AML	38(27/11)
Lung cancer	12533	MPM/ADCA	32(16/16)
DLBCL	6817	DLBCL/FL	77(58/19)

Table 2 Parameter values used

Parameter	Value
No. of generations (G)	200
Mutation probability (m_p)	0.15
Crossover probability (c_p)	0.9
Population size (N)	100
Mutation constant (C)	10

Table 3 Statistical measures of the population after final generation for a run

Dataset	Min.	Max.	Avg.	Std.
Leukaemia	0.9999	0.9999	0.9999	0.0000
Lung	0.9999	0.9999	0.9999	0.0000
DLBCL	0.8995	0.9451	0.9165	0.1356

Table 4 Experimental results with five different runs

Dataset	Run#	#Genes	Acc. (%)	FV value
Leukaemia	1.	5	100	0.9999
	2.	5	100	0.9999
	3.	5	100	0.9999
	4.	5	100	0.9999
	5.	5	100	0.9999
Lung	1.	4	100	0.9999
	2.	4	100	0.9999
	3.	4	100	0.9999
	4.	4	100	0.9999
	5.	4	100	0.9999
DLBCL	1.	6	92.2	0.9451
	2.	6	92.2	0.9451
	3.	6	92.2	0.9451
	4.	6	90.4	0.9143
	5.	7	85.7	0.8995

Table 5 Classification accuracy (%) using with and without proximity mutation

Dataset	Proximity mutation		Without proximity mutation	
	Best acc.	Avg. acc.	Best acc.	Avg. acc.
Leukaemia	100	100	100	100
Lung	100	100	97.36	96.71
DLBCL	96.1	94.475	94.8	94.15

A chromosome with the least number of genes, providing maximum classification accuracy is treated as the best gene for classification between cancerous and non-cancerous classes. The experiment is independently conducted several times on each dataset to evaluate the chromosomes using Eq. (1) with $\alpha = 0.7$ and $\beta = 0.3$, set experimentally. Table 4 shows the values of the fitness function for the chromosome with number of genes and corresponding accuracy for five such runs.

The gene selection method uses proximity mutation to maintain the diversity in GA environment. The Table 5 represents the results of the proposed methodology using proximity mutation and without proximity mutation. It is observed (from Table 5) that the using proximity mutation the proposed method gives better results than without proximity mutation which gives the superiority of the proposed mutation method. The Table 5 also shows the average classification accuracy using ten individual runs. From the result, it is observed that the best as well as average accuracies are same for Leukaemia dataset but for Lung and DLBCL dataset the proximity mutation gives better results.

Table 6 Comparative analysis between proposed and other methods described in literature

Dataset	Method	#Genes	Classification method	Accuracy (%)
Leukaemia	Hyk Gene [9]	4	KNN	98.61
	GA [12]	100	Fuzzy	98.61
	New-GASVM [11]	40	SVM	100
	α-value [10]	1–100	NB	100
	Proposed	5	SVM	100
Lung	α-value [10]	1–100	NB	100
	MFMW [13]	6	C4.5	98.34
	Proposed	4	SVM	100
DLBCL	α-value [10]	1–100	SVM/KNN/DT/NB	84.48
	Proposed	6	SVM	100

3.3 Comparison Analysis

In order to realize the goodness of the proposed system, a comparative analysis between the proposed method and the methods described in the literature [9–12] is performed, the results of which are shown in Table 6. The results reveal the fact that the proposed algorithm possesses the capability of finding highly significant genes and achieving relatively better classification performance than other methods. The results for existing methods are collected from corresponding papers where classifier with the maximum accuracy is listed in Table 6. Our method winds up taking a smaller number of active genes and giving higher classification accuracy (using SVM classifier) on a reduced dataset than the other methods.

4 Conclusion

Cancer classification is an important aspect of the disease and drug discovery problem using systematic and unbiased approach. Biologists need a small subset of genes to work with i.e. those genes which are primarily responsible for the cancer before they dive into elaborate experiments on the high volume microarray datasets which can be both time consuming and expensive. Hence, automated detection of this small and informative subset is highly profitable. In this paper, a novel proximity mutation based genetic algorithm has been proposed for selecting minimum number of relevant genes for cancer classification. Using GA with the proposed proximity mutation technique, the results of gene selection are better than conventional implementations of the algorithm. Our work is perhaps the first to use proximity mutation which yields better results than without such mutation related GA methods. The algorithm proposed is analyzed against existing state of the art

gene selection methods and run on publicly available microarray datasets in order to establish its efficiency and pave the way for further work on improving it still where scope remain be.

References

1. Mansouri, J., Khademi, M.: Multiplicative distance: a method to alleviate distance instability for high-dimensional data. Knowl. Inf. Syst. **45**(3), 783–805 (2015)
2. Kossenkov, A.V., Ochs, M.F.: Matrix factorization methods applied in microarray data analysis. Int. J. Data Mining Bioinform. **4**(1), 72–90 (2010)
3. Dudoit, S., Fridlyand, J., Speed, T.: Comparison of discrimination methods for the classification of tumours using gene expression data. JASA **97**, 77–87 (2002)
4. Goldberg, D.E.: Genetic Algorithms in Search, Optimization, and Machine Learning, p. 432. Addison-Wesley (1989)
5. Beasley, D., Bull, D.R., Martin, R.R.: An overview of genetic algorithms: Part 2 research topics. Univ. Comput. **15**, 170–181 (1993)
6. Hashem, A.-T., Alex, P.A.: Using genetic algorithms to solve optimization problems in construction engineering. Constr. Arch. Manage. **6**(2), 121–132 (1999)
7. Sung, H.J.: World Acad. Sci. Eng. Technol. Int. J. Comput. Electr. Autom. Control Inf. Eng. **3**, 8 (2009)
8. Shutao, L., Xixian, W., Xiaoyan, H.: Soft. Comput. **12**(7), 693 (2008)
9. Wang, Y., Makedon, F.S., Ford, J.C., Pearlman, J.: HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data. Bioinformatics **21**(8), 1530–1537 (2005)
10. Wang, X., Gotoh, O.: A robust gene selection method for microarray-based cancer classification. Cancer Inform. **9**, 15–30 (2010)
11. Mohamad, M.S., Deris, S.: A hybrid of genetic algorithm and support vector machine for features selection and classification of gene expression microarray. Int. J. Comput. Intell. Appl. **5**(1), 91–107 (2005)
12. Schaefer, G.: Data mining of gene expression data by fuzzy and hybrid fuzzy methods. IEEE Trans. Inf. Technol. Biomed. **14**(1), 23–29 (2010)
13. Leung, Y., Hung, Y.: A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification. IEEE/ACM Trans. Comput. Biol. Bioinf. **7**(1), 108–117 (2010)
14. Kent Ridge Bio-medical Data Set Repository. http://datam.i2r.a-star.edu.sg/datasets/krbd

Perspective Approach Towards Business Intelligence Framework in Healthcare

Mittal Kavita, S. K. Dubey and B. K. Sharma

Abstract Healthcare is highly complex industry driven by knowledge with rising cost and increasing demands for healthcare quality services. Healthcare providers are forced to focus on care quality while minimizing the cost through better healthcare resource management. However the abundant data from different sources such as clinical processes, business processes, and operational processes, causing remarkable issues and challenges are not resolved, through traditional technologies. Thus the Healthcare providers in effort to improve care quality and reduce cost are turning towards advanced and flexible IT-enabled business strategies. This paper attempts to illustrate the BI approaches incorporated with data mining techniques appropriate in the healthcare domain to overcome the issues and challenges more efficiently. Here emphasis is given on the main BI healthcare processes, benefits of using BI strategies in terms of efficiency, care quality and patient satisfaction.

Keywords Business Intelligence · BI Healthcare processes · BI Solutions
BI adoption · BI framework · Data mining

M. Kavita (✉)
Amity Institute of Information Technology, Amity University, Uttar Pradesh,
Sec-125, Noida, UP, India
e-mail: kavitamittal.it@gmail.com

S. K. Dubey
Amity School of Engineering and Technology, Amity University, Uttar Pradesh,
Sec-125, Noida, UP, India
e-mail: skdubey1@amity.edu

B. K. Sharma
Computer Science & Engineering Division, Software Development Centre,
Northern India Textile Research Association, Ghaziabad, India
e-mail: drbkjpr@ymail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_40

1 Introduction

The Healthcare industry being complex and knowledge based is more concerned about the quality of care with lesser cost. The fundamental need of health care is to achieve high care quality at lowered cost. The health care organization are forced to provide access to data from multiple domains such as clinical, financial, quality, and patient experience information [1]. This data needs to be integrated to clearly define the relationship between outcome and the cost [2, 3]. Healthcare providers in effort to improve care quality and reduce cost, are increasingly turning to advanced IT-enabled business strategies. For many years Business Intelligence has been adopted in large enterprises, particularly from the business sector and the banking sector. However, Healthcare industry is still deprived of flexible approaches of Business Intelligence technology that can be proved as a valuable tool to overcome the healthcare issues in terms of data integration, service quality and efficiency. Hence the research on the possibilities of using Business Intelligence in Healthcare has been strengthened but BI being a relatively new idea to healthcare is a biggest challenge. The need for a more systematic and deliberate study on Business Intelligence and the factors that allow for success in BI initiatives in healthcare organizations is crucial [4]. In terms of tools, general BI tools are used in healthcare industry, but people experienced in both healthcare data and Business Intelligence can be helpful in building custom data models to suit the needs of healthcare organization [5]. This paper focuses on how the issues in Healthcare can be targeted and overcome by more efficient and flexible BI approaches and understands the Business Intelligence Healthcare processes.

This paper is structured as follows. The Sect. 2 introduces research methodology. Section 3, presents the literature review. Section 4, perform the analysis, and Sect. 5 presents the evaluation of research questions. Finally, the paper is concluded.

2 Research Methodology

The base of Methodology for Research adopted here the literature review conducted on the relevant researches. The objective of this paper is to present a cognitive review related to the research area. The research methodology included all journals and research papers relevant to the field of research and is examined at different levels to target the objective. While planning the review methodology the requirements for writing the review are elucidated, and then the appropriate research questions are framed. While conducting the literature review, the initial researches are identified, then data is abstracted, analyzed and evaluated for its relevance and review report is presented. This review considered the research papers, articles, reports from 2002 to 2016 (till date) based on the availability of more research work during these years.

2.1 Research Questions on BI Framework in Healthcare

Two research question related to the problem were framed based on the reviewed literature.

RQ1: How can a Healthcare organization use the power of BI to overcome the issues and challenges in Healthcare?

RQ2: How can a BI set up be developed that provides an impact across the organization?

These research questions help to find the research gaps in previous and present research work related to Business Intelligence in Healthcare The review is conducted in more conventional and systematic manner by considering relevant researches related to the technical support, working papers, articles and Ph.D. thesis.

2.2 The Strategy for Primary Search

Digital libraries, white papers and other online articles were considered as a source of primary search. In the initial stage, various keywords related to Healthcare, Business Intelligence, BI frameworks were used frequently for searching relevant material.

2.3 Inclusion and Exclusion Strategies

While conducting the review process, the papers from 2002 to 2016 (till date) were included based on the availability of relevant work. To identify the required material from the study, notes were prepared from each paper. These records include title, abstract and findings from the study. Then at the next stage, only those notes were included which were relevant to our study. In this review process 31 research papers were selected relevant to the specified research work. Before 2002 not much significant work is discovered and however, no work is discovered in 2016 till date.

3 Literature Review

Business Intelligence in healthcare has become a subject of interest to the researcher due to rising demand of improvement in healthcare service quality, patient satisfaction and safety and organizational efficiency. The stakeholders of healthcare

support the adoption of advanced technology to enhance the service quality, availability of real time information and support to economic activities [6].

3.1 The Significance of Business Intelligence

BI is a wide combination of technologies, applications and processes for integration and analysis of data being beneficial to the stakeholders of healthcare organizations in making effective decision support system [7, 8]. BI has been considered to be valuable tool by many business organizations to reach their strategic goals, increase profitability and improve customer satisfaction [9]. The role of business intelligence is to deliver right information at right time and right location to improve the decision support process [10].

3.2 Business Intelligence in Healthcare

Discovering actionable information from huge amount of data is a complex task faced by healthcare providers today. A healthcare organization should be aimed at treating patient up to their level of satisfaction as well as achieving desired management outcomes. Two different approaches exist for Business Intelligence in Healthcare: data centric approach and process centric approach [11]. The data-centric approach combine operational data with OLAP tools to achieve effective decision making support by improving the quality of inputs to the decision process at reduced access time [12] which allows the firm to better understand its own capabilities [13]. The process-centric approach focuses more on the organizational processes helping in understanding the organizational capabilities through the integration of discovered knowledge with the organizational processes. The healthcare organization need to enhance its organizational capacity, standardization of business processes and improvement in the patient's treatment and care quality by implementing some effective solutions based on BI technologies [14].

3.3 BI Approaches

So, far we have been discussing about the traditional BI approach that has been struggling since a decade to satisfy the needs of Healthcare sector. However, the healthcare sector has gained lot of benefits by the adoption of BI technology but due to dynamically changing needs and complexities, the requirement arise for more efficient and flexible BI approaches among which the most commonly used are Cloud based BI and Mobile based BI. Cloud based BI applications are hosted by virtual network i.e., internet. It provides easy access, with less administrative tasks

related to data management and is scalable. Cloud based BI can perform all the functions provided by traditional BI more efficiently and lesser cost [15].

Mobile based BI is advancement over traditional and cloud BI that provide access to BI related data on Mobile devices. Mobile BI is capable of handling the use case of mobile users that need remote access to critical business, clinical and financial information. The developers of smart phones have provided the platform for development of mobile based BI applications [16]. Data collected by organizations can be converted into useful knowledge due to the use of advanced data warehousing and analytics tools [17]. The difficulties arising in adoption of Mobile BI adoption today are: Lack of technical skills, complexity of mobile based IT systems, future uncertainties, rising cost in technology transformation, patient privacy. Moreover the Mobile BI is more beneficial from patients view point enabling the patients to: enhance awareness and participation in care, improved access to preventive healthcare information, enhance communication with healthcare providers, reduction in healthcare delivery cost [18].

3.4 Available BI Solutions and Their Adoption

In healthcare organizations, the factors responsible for organization performance and efficiency are changing rapidly with the changes in the needs of healthcare sector and advancement in technology [19]. Business Intelligence is capable of providing one solution to multiple problems related to managing data, quality, handling organizational processes, monitoring performance, and many others. BI solutions are tools for analysis and monitoring of organizational performance [20]. There are many BI solutions developed with their unique functionality and technology. As per the Business intelligence market analysis report the major contribution is in North America by U.S., followed by Asian countries including India, China, Singapore, Malaysia, etc., the European countries are less competitive due to some barriers to BI adoption [21]. The business intelligence solutions developed for healthcare that are performing well in terms of managing and integrating data, interoperability of processes, reporting and visualization, performance management [18]. The organizations adopting BI are gaining benefits such as: improved performance, increased analytical and visualization capabilities, effective and flexible business processes with efficient decision making [22, 23].

3.5 BI Maturity Models and Frameworks

Maturity Models provide a systematic framework to identify the strengths and weaknesses of an organization and enable continuous enhancement in the efficiency of an organization to achieve strategic goals [24–26]. BI maturity Models provide strategic guidelines to enhance efficiency and assessment criteria for data standards

and quality services [27–29]. A number of maturity models exist in the literature which are targeting Business sector, banking sector, enterprises, but none of them are targeting to the current needs of healthcare sector. Hospitals can adopt BI systems to enhance the care quality, patient satisfaction and operational efficiency in terms of medical and business [30]. Clinical Research using BI framework can optimize the transformation of data into knowledge and clinical research process [31].

3.6 Data Mining for Business Intelligence in Healthcare

Data mining technology introduces a variety of techniques to determine relationship among data, finding hidden patterns in huge data, make interpretations to predict future trends and support effective decision making. The data mining techniques being elements of statistics, machine learning, regression models; help organizations to understand and identify trends within the huge data [32]. The algorithms can be applied with their suitability depending on the application domain [33].

4 Analysis

4.1 Healthcare BI Processes

Healthcare organizations are composed of a variety of integrated processes that needs operational support. In relation to BI, Healthcare organizational processes can be differentiated in three categories [34]: Clinical Processes, Business Processes, and Operational Processes (Fig. 1).

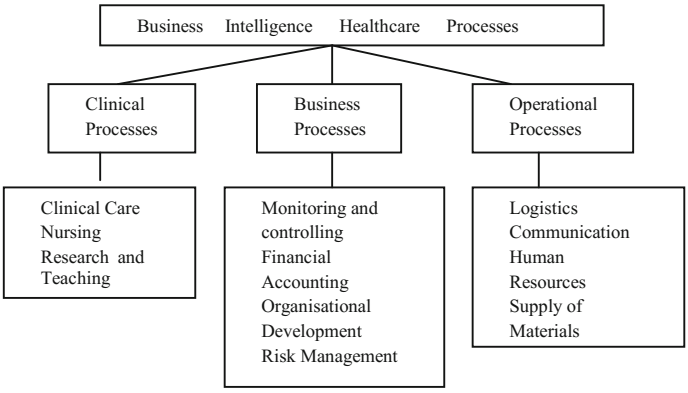


Fig. 1 Healthcare BI processes

4.2 BI Approaches: Comparison

BI technology is used now a days in more advanced forms to meet the needs of current healthcare sector. The use of BI at initial stage began with traditional BI solutions that are powerful and supported by well established software companies. But using traditional BI systems require high level technical skills, IT infrastructure equipped with SQL servers that pushes up the cost and time consuming deployment of the solution. Also in spite of being powerful these systems are static, based on historical data and trained professionals for their effective use. Since many years, cloud BI is developing its importance as an alternative to traditional BI solutions. The cloud BI is in demand and in use due to its easy to use features. But now days due to advancement in mobile computing and mobile devices there is a need for lightweight version of BI i.e. Mobile BI. Healthcare organizations are looking for adoption of Mobile BI for real time analytics, better visualization with affordable resources at lower cost.

4.3 Healthcare Issues and Challenges

The healthcare industry is forced to meet the needs of discovering real time information from huge varied data sources. However, on the basis of review, in terms of BI implementation the healthcare sector face the following challenges:

- Challenges for quick access to integrated information.
- Challenges for advanced and complex mobile technology.
- Challenges for care and service quality.
- Challenges for patient satisfaction and safety.

For BI to be successful in healthcare sector, it is necessary to understand the complexities and challenges of Healthcare and how BI can be impacted.

4.4 BI Solutions—Analysis

The needs of healthcare organizations have changed during the years and became more complex in competitive environment. There are a number of BI solutions exist that began using traditional BI approach to overcome the needs in healthcare sector. The BI solutions are performing well in terms of analytics, reporting, data warehousing, and performance management focusing on all financial, clinical and operational process of an organization but there is less consideration towards care service quality, patient satisfaction and safety. Due to technical complexities in Traditional BI, the providers are turning towards the alternative flexible approaches. Many of BI solutions are using cloud based BI as an alternative to traditional BI but only few have incorporated the Mobile BI to deal with wireless world and smart phones.

5 Result

This paper is intended towards analyzing the challenges in Healthcare industry and how these challenges can be overcome by using intelligent mobile based BI framework in healthcare. This section shows evaluation of solution of the research questions framed.

RQ1: How can a Healthcare organization use the power of BI to overcome the issues and challenges in Healthcare?

After performing the analysis of various researches published, it can be indicated that the research in this area is not satisfactory in terms of healthcare needs and thus needs more efforts. From 2002 to 2016 the related research works on Healthcare needs targeted by BI is shown in Fig. 2. Graph indicates the research work in related area so far in 15 years.

RQ2: How can a BI set up be developed that provides an impact across the organization?

While going through various research papers for BI frameworks or models developed during the 15 years, it can be analyzed that there is sufficient use of BI and its approaches for resolving issues in healthcare some are using traditional BI approach, cloud BI, Mobile BI. From the Fig. 3, it can be deducted that only few BI solutions in healthcare has adopted Mobile BI. Thus, future research can be carried with Mobile BI platform to resolve the issues in healthcare sector with more benefits at lesser cost.

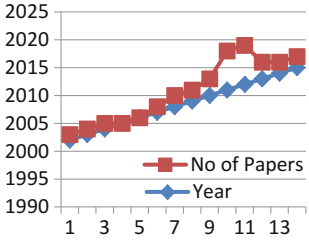


Fig. 2 Literature review analysis (2002–2016)

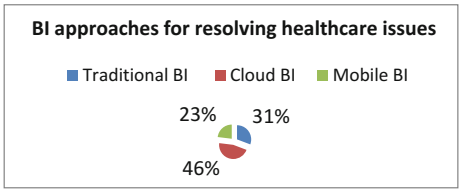


Fig. 3 BI approaches in healthcare

6 Conclusion

After analyzing the review, it can be stated that healthcare organizations are achieving a lot of benefits using Business Intelligence systems. Business Intelligence systems provide reliable and consistent information from all the areas of organization activity. Nowadays, BI systems have been proved to be crucial in developing effective decision support systems that will help improve patient's clinical outcomes and the quality of medical services. But the traditional BI being more rigid and complex, there is need to move more advanced and flexible BI approaches with more focus towards mobile BI. From the review it can be analyzed that Mobile BI is performing better to satisfy the need for higher efficiency by accessing real time data quickly and at less cost for making real time decisions. Moreover, BI solutions in India are more focused towards performance monitoring functionalities but less diverted to person needs. This research will be more directed towards patient oriented framework focused towards care quality, patient satisfaction, and information quality. By adopting proposed BI framework, healthcare efficiency can be enhanced by extracting valuable information remotely, with improved service quality, patient satisfaction and safety.

References

1. Microsoft, Business Intelligence for Healthcare: the New Prescription for Boosting Cost Management, Productivity and Medical Outcomes, an exclusive report from Business Week Research Services (2009)
2. Doyle, M.: Getting the best business intelligence solution for healthcare. Whitepaper, Healthcatalyst (2013)
3. Houghton, J.: Information technology and the revolution in healthcare. In: Equity, Sustainability and Industry Development Working Paper Series (2002)
4. Olszak, C.M., Ziemba, E.: Approach to building and implementing business intelligence systems. Interdiscip. J. Inf. Knowl. Manage. **1**(2), 135–148 (2007)
5. Briggs, L.L.: https://tdwi.org/articles/2013/06/25/Healthcare-BI-Challenges-Opportunities.aspx?share. Accessed 22 Nov 2015
6. Ashrafi, N., Kelleher, L., Kuilboer, J.-P.: The impact of business intelligence on healthcare delivery in the USA. Interdiscip. J. Inf. Knowl. Manage. **9**, 117–130 (2014)
7. Kolowitz, B.J., Shresth, R.B.: Enabling business intelligence, knowledge management and clinical workflow with single view. Issues Inf. Syst. **12**(1), 70–77 (2011)
8. Wixom, B., Watson, H.: The BI-based organization. Int. J. Bus. Intell. Res. **1**(1), 13–28 (2010)
9. Olszak, C.M., Batko, K.: Business intelligence systems-new chances and possibilities for healthcare organizations. Bus. Inform. J. **3**(25), 123–138 (2012)
10. Koh, H.C., Tan, G.: Data mining applications in healthcare. J. Healthcare Inf. Manage. **19**(2), 64–72 (2011)
11. Mettler, T., Vimarlund, V.: Understanding business intelligence in the context of healthcare. Health Inform. J. **15**(3), 254 (2009)
12. Magda, I., Szczygielski, K.: An Assessment of Possible Improvements to the Functioning of the Polish Healthcare System (2012)

13. Negash, S.: Business intelligence. Commun. Assoc. Inf. Syst. **13**(1), 15 (2004) (ebook)
14. Olszak, C.M., Batko, K.: The use of business intelligence systems in healthcare organisations in Poland. In: Proceedings of the Federated Conference on Computer Science and Information Systems, pp. 969–976 (2012)
15. https://www.klipfolio.com/resources/articles/what-is-cloud-business-intelligence. Accessed 17 Apr 16
16. https://www.microstrategy.com/Strategy/media/downloads/solutions/MicroStrategy-Mobile-Healthcare-Providers-Brochure.pdf. Accessed 17 Apr 16
17. Watson, H.J.: BI-based organizations. Bus. Intell. J. **15**(2) (2010)
18. Power to the patient: How mobile technology is transforming healthcare, A report from The Economist Intelligence Unit, SAP, 2014. http://www.economistinsights.com/sites/default/files/HowMobileisTransformingHealthcare.pdf. Accessed 17 Apr 16
19. Bogdan, A., Sorina, P.: Business intelligence. A presentation of the current lead solutions and a comparative analysis of the main providers. Database Syst. J. **V**(2) (2014)
20. Alexandra, R.: Comparative analysis of the main business intelligence solutions. Informatica Economică **17**(2) (2013)
21. http://www.marketsandmarkets.com/Market-Reports/healthcare-business-intelligence-market-252368925.html. Accessed 17 Apr 16
22. Leonardi, T.: Business Intelligence and Healthcare. The Cornerstone of Any Successful Healthcare Organization Will Be "Healthcare Business Intelligence (2008)
23. Microsoft, Knowledge Driven Health. Think Bigger about Business Intelligence—Create an Informed Healthcare (2012)
24. Lahrmann, G., Marx, F., Winter, R., Wortmann, F.: Business intelligence maturity: development and evaluation of a theoretical model. In: 44th Hawaii International Conference on System Sciences, Kauai, Hawaii (2011)
25. Brooks, G.P., El-Gayar, O., Sarnikar, S.: Towards a business intelligence maturity model for healthcare. In: 46th Hawaii International Conference on System Sciences, Hawaii (2013)
26. Glancy, F.H., Yadav, S.B.: Business intelligence conceptual model. Int. J. Bus. Intell. Res. **2** (2), 48–66 (2011)
27. Gaddum, A.: Business Intelligence (BI) for Healthcare Organizations (2012)
28. Tavallae, R., Shokohyar, S.: Assessing the evaluation models of business intelligence maturity and presenting an optimized model. Int. J. Manage. Account. Econ. **2**(9) (2015). ISSN 2383-2126
29. Tan, C., Sim, Y.W., Yeoh, W.: A maturity model of enterprise business intelligence. Commun. IBIMA Article ID 417812 (2011)
30. Kao, H.-Y., et al.: Implementing BI to assist decision making in healthcare: a case of Regional Taiwanese Hospital. In: International Conference of European Federation for Medical Informatics (2012)
31. Keeling, T.L.: Clinical research: using business intelligence framework. J. Issues Inf. Syst. **XI** (1), 372–376 (2010)
32. Diwani, S., Sam, A.: Framework for data mining in healthcare information system in developing countries: a case of Tazania. Int. J. Comput. Eng. Res. **03**(10) (2013)
33. Kayange, D.S.: Overview applications of data mining in health care: the case study of Arusha region. Int. J. Comput. Eng. Res. **03**(8) (2013)
34. Katoua, H.S.: The benefits of using data mining approach in business intelligence for healthcare organizations. Egypt. Comput. Sci. J. **36**(2) (2012)

Gene Selection and Enrichment for Microarray Data—A Comparative Network Based Approach

Debasish Swapnesh Kumar Nayak, Saswati Mahapatra and Tripti Swarnkar

Abstract Gene selection plays a vital role in understanding the disease progression and further it helps in understanding the therapeutic targets. Most of the genes available in micro array data are not informative for a particular disease of interest. Study of functional analysis and interaction structure of genes plays a vital role in selecting genes associated to complex diseases. This work uses two different network based approaches for gene selection and compares the biological and statistical enrichment of selected genes. Functional modules in the gene expression data are obtained using Gene Correlation Network (GCN) and marker genes in the modules are identified using R package Weighted Gene Co- expression Analysis (WGCNA). WGCNA is considered to be one of the best methods for analysis of global GCN using a suitable threshold that leads to a network with scale free topology. The differentially co-expressed genes are then compared with the existing gene selection approach which integrates the selected co-expressed gene modules with protein-protein interaction (PPI) network. Observation shows that using PPI network which is generated using multitude of high throughput experiments and available in public data bases selects more disease specific genes in comparison to constructed GCN. The study shows that integrative network analysis to find genes may provide greater insight in underlying biological response.

Keywords Gene correlation network (GCN) · Protein protein interaction network (PPI) · Gene selection · Gene enrichment · Biological network integration · Weighted gene correlation network analysis (WGCNA)

D. S. K. Nayak (✉)
Indian Institute of Technology, Bhubaneswar, India
e-mail: swapnesh.nayak@gmail.com

D. S. K. Nayak · S. Mahapatra · T. Swarnkar
Department of Computer Application, Siksha 'O' Anusandhan University, Bhubaneswar, India
e-mail: saswatimohapatra@soauniversity.ac.in

T. Swarnkar
e-mail: triptiswarnakar@soauniversity.ac.in

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_41

1 Introduction

DNA micro array represents the state of a cell at a molecular level and has the capability to analyze thousands of genes in a single experiment. Microarray is a high throughput gene expression data simultaneously monitoring thousands of genes. The study of microarray data is limited by high dimension of features or genes with comparatively less number of samples. The small sample size becomes a limitation in various analysis [1]. Thus, gene selection plays an important role in analysing gene expression and helps to identify the candidate genes that can be further analyzed for disease prognosis [2]. Gene co-expression network which is the collection of co-expressed genes have been found successful to describe the pair wise relationship between gene transcripts [3]. GCN is used to identify modules of genes with similar expression profiles. Apart from focusing on restrictive single data analysis, integrative analysis of biological data at different level provides more reliable and complete information about the genotype as well as the phenotype association. Protein-protein interaction (PPI) combines structural biology and bioinformatics to find the physical interactions among the pair of proteins. Genes that are related to some specific disease need not to be differentially expressed, but may play important role in interconnecting the differentially expressed genes in the PPI network [4].

In this work we have compared two network based approaches for gene selection. Gene co expression network (GCN) is constructed using the tool weighted gene co-expression network analysis and the functional modules in the network are identified. Genes are selected from the modules by ranking. The efficiency of selected genes is then compared with the existing method which integrates co-expression gene clusters with PPI network for selecting the marker genes.

2 Related Work

Individual analysis of biological data at multiple omic level results in incomplete understanding of genetic aetiology of the complex traits. Intergeneration of multiple omic data is expected to compensate for any undependable or noise information in any single data type and is unlikely to lead to false positives [5]. Swarnkar et al. proposed an integrated framework that combines gene expression information with structural facts of PPI networks to identify a set of functionally enriched genes associated with a specific disease [1]. They have identified co-expressed gene modules in the gene expression data set using k-means clustering algorithm which are mapped to PPI network available in standard public data bases in order to find dense sub graph (DSG) in the network.

Biological network gives valuable information in studying system level properties. It can give better insight on disease progression via the identification of perturbed set of genes in different complex diseases. Network biology approach

uncovers the underlying mechanisms in disease pathogenesis, identification of new biomarkers, and shed light on personalized therapeutic interventions [6].

WGCNA, a R package [3] for gene co-expression network analysis has been proven to be a well accepted method for global analysis of co-expressed genes and modules. The package provides different R functions environment to study the various aspects of weighted correlation network analysis. Kadarmideen et al. compared two different methods WGCNA and partial code information theory (PCIT) for GCN construction and analysis. They found that WGCNA method is favourable over PCIT method as WGCNA retains biologically relevant hub genes and their connections within sub-networks intact where as PCIT deletes some important edges in the network and hence disrupts the network topology [7].

3 Materials and Methods Used

3.1 GCN

Gene co-expression network (GCN) considered for graphical representation of genes, where each node of the graph is represented as a gene and a pair of nodes is connected with an undirected edge. An undirected edge is found between a pair of genes only if it's pair-wise expression similarity is above a particular threshold. Construction of co-expression network using the gene expression information is considered as one of the best alternative to the traditional analysis approaches [2]. Large-scale gene co-expression networks analysis shows that the biologically related genes are highly co-expressed across different organisms and across multiple datasets. In GCN, nodes represent genes where node profile x_i represents gene expression profile. The Gene Correlation Network is mainly represented using adjacency matrix a_{ij} which is constructed considering the co-expression similarity between genes i and j.

3.2 WGCNA

To study the various aspects of weighted correlation network analysis, R provides Weighted Gene Correlation Network Analysis (WGCNA) a software package which comprehends a large collection of R function is being widely used in literature [3].

WGCNA can be used to construct the highly co-expressed gene modules as co-expressed gene cluster from the given gene expression data. Further analysis of these co-expressed gene modules using WGCNA may provide us insight about the representative gene or an eigen gene or an inter modular hub gene in each module. Further it can also provide insight about inter modular connectivity, their relation

with external traits and for calculating module membership measures. This analysis of gene correlation network gives further insight in finding candidate biomarkers as therapeutic targets. The approach has been used in various biological context, viz., yeast and mouse genetics, cancer, brain image data analysis etc. [2].

3.3 Datasets

DNA microarray data set for homo-sapiens available in NCBI's Gene Expression Omnibus (GEO) has been used in our study. We have used prostate cancer gene expression data set for gene selection process. Prostate cancer data consists of 20,000 genes and 104 samples, out of which 34 are normal samples and 70, are cancer related samples. To compare the quality of modules formed using GCN approach with the existing PPI network based approach, we have used cancer gene data set available in NCBI. The cancer data set contains 10807 genes and 17 samples.

3.4 Working Model

Figure 1 represents the work flow of the model being used to construct the gene correlation network (GCN) from the microarray gene expression data.

1. We have followed the pre-processing method as described in the existing method [1]. The normalized data from NCBI is collected and missing values in the data are removed by interpolating them using mean or median. Finally the

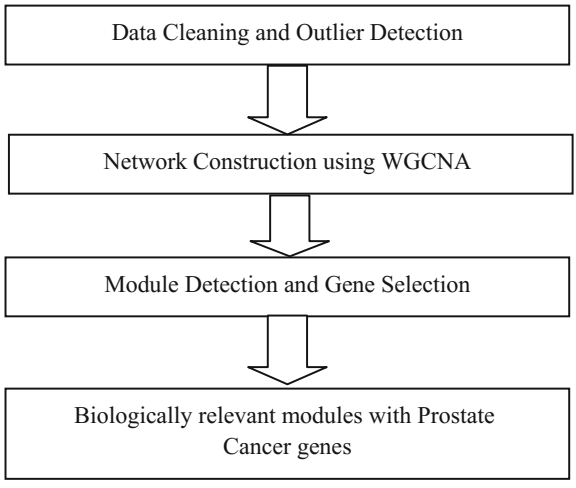


Fig. 1 Steps of WGCNA based method for gene selection and enrichment

variance across the samples is used for filtering genes from the given dataset. Thus, the size of the data set is reduced to 13,791 genes from 21000 genes and the samples remaining unchanged with 104. Using WGCNA function samples are clustered to identify outlier in the samples. After data cleaning and removing the outlier in the samples we have found 100 samples that are used for further analysis. Figure 2 shows the result of sample clustering.

2. In GCN modules in the network corresponds to cluster of genes with high absolute co-relation. To construct the network out of the pre-processed data and identify the modules in the network, we have used WGCNA function blockwiseModules () with soft-thresholding power = 4 and taking minimum module size to be 50. Figure 3 represents the scale free topology, as well as the mean connectivity for different soft thresholding powers.

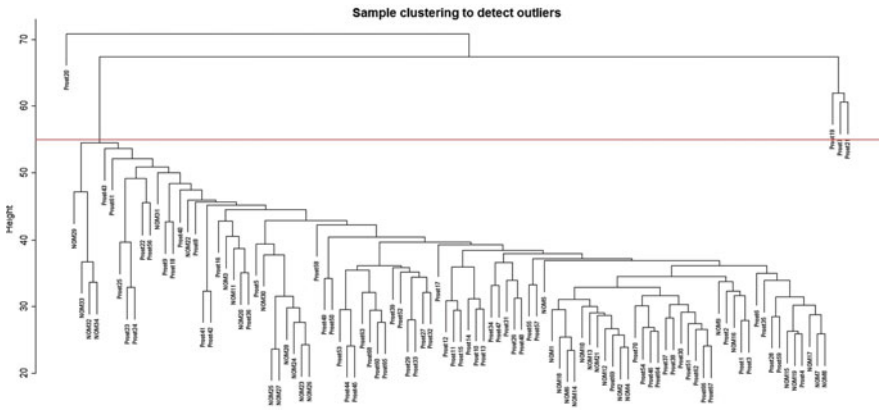


Fig. 2 Sample clustering with cut height chosen as 55

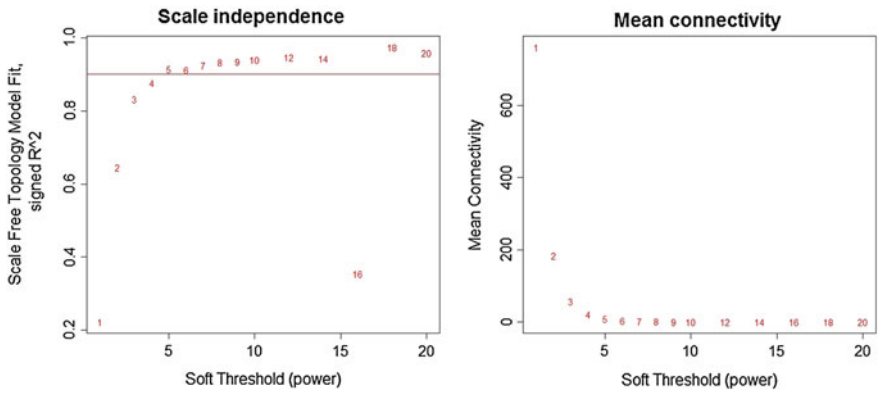


Fig. 3 Scale free topology to find out the soft threshold power

3. The eigen genes which are the representative genes in each module are identified. Module membership values for all genes with respect to the Eigen genes in the modules are calculated and genes in the modules are ranked according to their module membership values and few top ranked genes are selected. Effectiveness of subset of the selected genes are computed by using different classification techniques.
4. In order to measure the quality of modules, we have compared the modules obtained using our GCN based approach with the modules formed in the existing PPI based approach for gene selection.

3.5 Performance Measures Used

The Matthews coefficient correlation (mcc) is used as a measure of quality of binary classification and is regarded as a balanced measure and can be used for the classes which are of very different sizes [8]. Prostate dataset is having imbalanced ratio between number of samples in positive and negative classes. Thus, to measure the predictive accuracy of the selected genes we are using the mcc. In mcc, overall accuracy, sensitivity, specificity precision and *f*-measure used for comparison to the known true classes are defined as follows.

$$mcc = \frac{tp * tn - fp * fn}{\sqrt{(tp + fp) \times (tp + fn) \times (tn + fp) \times (tn + fn)}} \tag{1}$$

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{2}$$

$$sensitivity = \frac{tp}{tp + fn} \tag{3}$$

$$specificity = \frac{tn}{tn + fp} \tag{4}$$

$$precision = \frac{tp}{tp + fp} \tag{5}$$

$$f - measure = \frac{2 \times tp}{2 \times tp + fp + fn} \tag{6}$$

where true-positive samples is denoted as tp, tn stands for the count of true-negative samples, fp represents the number of false-positive samples and fn is the number of false-negative samples. The above measures are being used for class performance analysis and comparison with existing methods in the literature. Samples are considered to be divided into two categories, namely diseased samples (positive)

and normal samples (negative). The comparative biological significance analysis of the modules obtained in the GCN based approach with the existing PPI network based approach is being made by studying the presence of disease related genes in each of these selected modules.

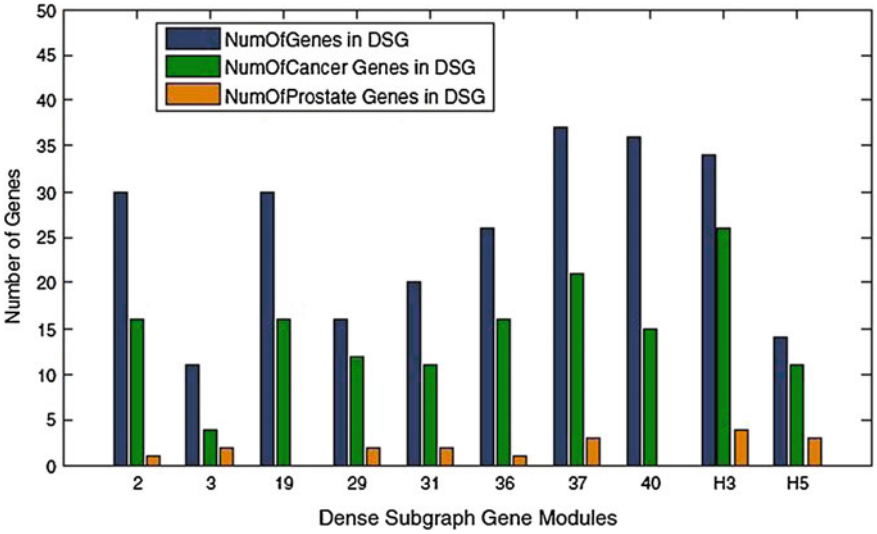


Fig. 4 Biological significance study of the PPI interaction approach gene modules in terms of the presence of disease-related genes [1]

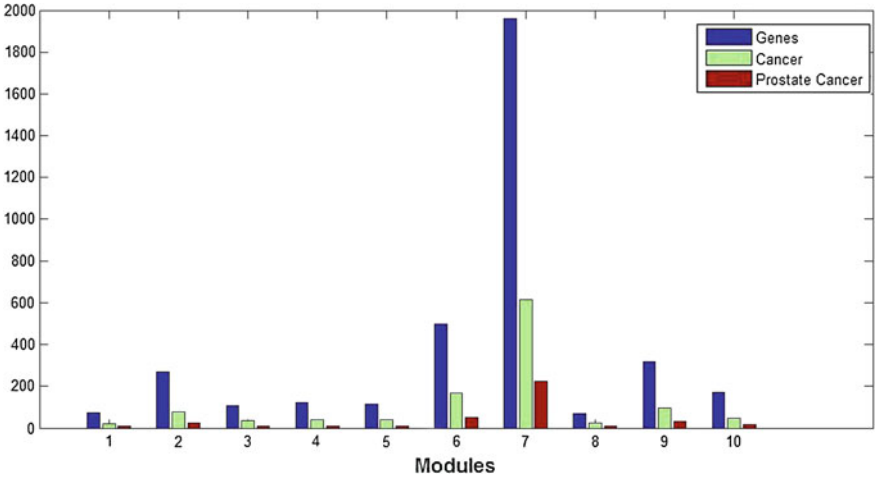


Fig. 5 GCN based module of biological significance in terms of presence of disease related genes

4 Results and Discussion

We have obtained 18 modules after network construction and module detection. The average module membership of genes in a module taken into consideration, we select 10 modules which are having highly correlated genes shown in the Fig. 5. The highly correlation leads to the higher quality measure. In Figs. 4 and 5 we have shown the percentage of prostate cancer related genes in modules obtained by PPI based gene selection approach and GCN based module selection approach, respectively.

In Table 1 we take top 10 modules of PPI based approach as well as from GCN based approach. The result states that the percentage of cancer genes with respect to the total number of genes in a module is high in case of PPI integration modules in

Table 1 Percentage of cancer and prostate cancer genes with respect to total genes in different modules. Where CG is Cancer Genes, PCG is Prostate Cancer Genes

Modules	PPI integrated gene modules			GCN gene modules		
	Genes	% of CG	% of PCG	Genes	% of CG	% of PCG
1	30	53.33	03.33	75	28.00	10.66
2	11	36.36	27.27	271	28.04	08.85
3	30	53.33	00.00	107	32.71	10.28
4	16	68.75	12.50	122	31.96	09.01
5	20	55.00	10.00	115	33.91	07.82
6	26	61.53	03.84	499	33.46	10.22
7	37	56.75	08.10	1962	31.29	11.41
8	36	41.66	00.00	68	36.76	11.76
9	35	71.42	11.42	320	30.00	09.68
10	13	84.61	23.07	172	27.90	09.88

Table 2 Percentage of prostate cancer genes with respect to cancer genes in different modules. Where CG is cancer genes, PG is protest genes

Mod	PPI integrated gene modules				GCN gene modules			
	Genes	CG	PG	% of PG	Genes	CG	PG	% of PG
1	30	16	01	06.25	75	21	08	38.09
2	11	04	03	75.00	271	76	24	31.57
3	30	16	00	00.00	107	35	11	31.42
4	16	11	02	18.18	122	39	11	28.20
5	20	11	02	18.18	115	39	09	23.07
6	26	16	01	06.25	499	167	51	30.53
7	37	21	03	14.28	1962	614	224	36.48
8	36	15	00	00.00	68	25	08	32.00
9	35	25	04	16.00	320	96	31	32.29
10	13	11	03	27.27	172	48	17	35.41

Table 3 Class performance of prostate cancer GCN based gene modules. mcc Matthews correlation coefficient, sen sensitivity, spec specificity, prec precision, fm f-measure, 3nn 3 nearest neighbours, rf random forest, svm support vector machine are expressed in percentage. M is Module, S is Size and C is Class

M	S	C	3nn					rf					svm				
			sen	spec	prec	fm	mcc	sen	spec	prec	fm	mcc	sen	spec	prec	fm	mcc
17	35	N	73	14	68	71	58	46	01	93	62	58	03	00	100	06	15
		P	85	26	88	87	58	98	53	81	89	58	100	96	70	82	15
8	50	N	90	21	64	75	63	56	02	89	69	62	10	00	100	18	26
		P	78	10	94	85	63	97	43	84	90	62	100	90	72	83	26
16	50	N	86	15	70	77	67	66	04	87	75	67	30	00	100	46	48
		P	84	13	93	88	67	95	33	87	91	67	100	70	76	87	48
13	50	N	66	15	64	65	50	46	02	87	60	54	00	00	00	00	00
		P	84	33	85	84	50	97	53	81	88	54	100	100	70	82	00
14	50	N	66	07	80	72	63	43	01	92	59	55	03	00	100	06	15
		P	92	33	86	89	63	98	56	80	88	55	100	96	70	82	15
4	50	N	70	20	60	64	48	43	07	72	54	43	00	00	00	00	00
		P	80	30	86	83	48	92	56	79	85	43	100	100	70	82	00
1	40	N	83	10	78	80	72	56	08	73	64	52	06	00	100	12	21
		P	90	16	92	91	72	91	43	83	87	52	100	93	71	83	21
18	50	N	83	18	65	73	61	53	01	94	68	63	03	00	100	06	15
		P	81	16	91	86	61	98	46	83	90	63	100	96	70	82	15
6	40	N	70	15	65	67	53	53	05	80	64	54	10	00	100	18	26
		P	84	30	86	85	53	94	46	82	88	54	100	90	72	83	26
10	40	N	63	11	70	66	53	43	02	86	57	51	00	00	00	00	00
		P	88	36	84	86	53	97	56	80	87	51	100	100	70	82	00

comparison to WGCNA based GCN approach the percentage of prostate cancer genes with respect to total number of genes in a module is more in few modules of PPI based approach and average in few modules of GCN based approach.

Table 2 shows the percentage of prostate specific genes with respect to cancer related genes in each selected modules. The result shows that except in module 2 the percentage of prostate cancer genes with respect to cancer genes is more in GCN based approach in comparison to PPI based approach. This shows that WGCNA is more efficient in finding co-expressed genes in comparison to cluster based approach used in PPI integration model.

Genes are ranked in decreasing order of their module membership values. On the basis of top few genes are selected for measuring class performance of each module. To evaluate the predictive performance of these genes, we have classified the genes using three different classifiers—K-Nearest Neighbour (knn for k = 3), Support Vector Machine (svm) and Random Forest (rf). The classifiers are applied with 10 fold cross validation. Table 3 summarizes the Sensitivity, Specificity, Precision, F- measure, mcc measures for each of the co-expressed prostate dataset for different classifiers. The result shows that highly correlated modules show higher value lower variance of mcc measure for different classifiers like k-nn, rf and svm. Thus, these selected modules 8, 16, 1 and 6 of GCN for prostate cancer can be further considered for statistical and biological in depth biological analysis. These modules may further give an insight in disease progression and may help in therapeutic analysis.

5 Conclusion and Future Work

It is observed that the network integration based gene modules are more significant in comparison to the traditional expression based gene selection. The study reveals that the module based network integration gene selection is able to find genes which are more discriminative and are found to play vital role in maintaining the interaction among the important genes.

The said property is important for the discovery of disease causing genes. The enrichment achieved by network integration using PPI is found to be stronger compared to that of the GCN based gene selection approach. In few biological analysis done by us it is observed that finding co-expressed gene modules (GCN) is more effective than cluster based approach used in compared PPI integration model. This states that the integration of PPI network with GCN may be studied further in future to find higher label interaction among multiple small co-expressed GCN modules. This may provide more accurate value to the pathway structures and will help in understanding network label biological dynamic in disease progression.

The genes selected using different level biological network integration approaches may be more relevant for the further study of progression of a specific disease. Both the compared approaches, viz., GCN based and PPI network based, can be integrated in different ways for more thorough analysis of candidate gene selection.

References

1. Swarnkar, T., et al.: Identifying dense subgraphs in protein–protein interaction network for gene selection from microarray data. Netw. Model. Anal. Health Inform. Bioinform. **4**(1), 1–18 (2015)
2. Singh, Rabindra Kumar, Sivabalakrishnan, M.: Feature selection of gene expression data for cancer classification: a review. Proc. Comput. Sci. **50**, 52–57 (2015)
3. Langfelder, P., Horvath, S.: WGCNA: an R package for weighted correlation network analysis. In: BMC Bioinformatics 9.1 (2008)
4. Chuang, H.Y., et al.: Network-based classification of breast cancer metastasis. Mol. Syst. Biol. **3**(1) (2007)
5. Ritchie, M.D., et al.: Methods of integrating data to uncover genotype-phenotype interactions. Nat. Rev. Genet. **1**6(2), 85–97 (2015)
6. Furlong, L.I.: Human diseases through the lens of network biology. Trends Genet. **29**, 150–159 (2013)
7. Kadarmideen, H.N., Watson-Haigh, N.S.: Building gene co-expression networks using transcriptomics data for systems biology investigations: comparison of methods using microarray data. Bioinformation **8**(18), 855–861 (2012)
8. Dao, P., et al.: Optimally discriminative subnetwork markers predict response to chemotherapy. Bioinformatics **27**(13), 205–213 (2011)

Part IV
Big Data and Recommendation Systems

Role of Big Data in Make in India

Sandeep Tayal, Nishant Nagwal and Kapil Sharma

Abstract Big Data is the buzzword in the field of technology for some time now. The demand for Big Data Technology is now felt by each and every organization of the world. The benefits of Big Data are immense. It is a study of role Big Data can play in Make in India. The make in India is about manufacturing products in India. The role Big Data can play in Manufacturing can revolutionize the entire process of manufacturing. Not only it can decrease manufacturing cost losses but can also help companies achieve customer satisfaction. From better streamlining of processes in a manufacturing unit to help create a better working environment, big data has brought a wave of change that cannot be ignored. Further, the Make in India not only brings a lot of investment but a lot of jobs too in India and we really in need of a technology to manage a large amount of data generated. By use of Big data, companies can implement better production techniques and thus can get an edge over their competitors. This study aims to help companies in India and abroad to understand the benefits of Big Data in manufacturing their products. It may also help companies who have already implemented Big Data but are not taking full advantage of it.

Keywords Big data paper · Product design · Manufacturing · Big data analytics Make in india · Machine learning · Production · Survey · Report

S. Tayal (✉) · N. Nagwal
Maharaja Agrasen Institute of Technology, Delhi, India
e-mail: tayal.mait@gmail.com

N. Nagwal
e-mail: nishantnagwal19@gmail.com

S. Tayal · K. Sharma
Delhi Technological University, Delhi, India
e-mail: kapil@ieee.org

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_42

1 Introduction

Make in India was launched by the Indian Government on 25 September 2014, so that the multi-national and the national companies can manufacture their products in India. The Make in India focuses on the creation of jobs and enhancement of skills in 25 sectors of the Indian economy. It also aims to bring the high-quality product standards which are further combined with minimum environmental impact. It also aims to bring technological as well as capital investment in India [1].

Big data is a buzzword which means a massive volume of data that is very large and it is difficult to process and analyze the data using normally used database and software techniques. The rate data is generated is too fast and the database becomes, too large for the traditional databases to handle [2].

Companies can improve their operations and, make faster and more intelligent decisions with the help of Big Data analysis.

Big Data analysis works in the favour of manufacturing industries on multiple fronts. From reducing product flaws to improving production quality while raising overall efficiency, not to mention saving a lot of time and money, the advantages are being garnered by many businesses around the world.

The applications of Big Data is in every field ranging from retail to the healthcare sector. All organization which keep an access to the Big Data can use it in analytics and thus can manage their operations. Big Data can play a very important role in helping companies develop new business models. Additionally, Big Data helps companies become proactive in cases of operational decision making in areas of improving product design, production failure, etc. In the field of manufacturing, the operations managers can use Big Data analytics to check data, and to identify various patterns and relationships among different small process steps and inputs, and thus can optimize the factors to increase product yield. Big data can help companies in improving the process of manufacturing by-product assessing the process interdependencies [3] using big data analytics and were able to identify the different parameters. It also has applications in the custom design on products [3]. It offers the high assurance of quality [3] and managing supply chain risk [3]. This, Big Data has a lot of roles when it comes to manufacturing and it can surely change the entire manufacturing process.

2 Big Data in Manufacturing

1. **It increases the accuracy, quality and the yield of the pharmaceutical production companies**. There are around 200 variables which need to be taken care of to ensure the purity of ingredients. The yields can vary from 50 to 100%. Based on Big Data analytics the yield was increased by around 50% [4].
2. **It helps in accelerating the IT, manufacturing and operational systems integration**. Big Data can be used for the optimization of the various production

schedules on different constraints. Big Data analytics will become a critical reason for the success of various multifunctional departments [4].

3. **It helps to integrate analytics over Six Sigma framework**. It helps in making the production workflows better by analyzing the various aspects of production and helps in delivering optimum quality products [4].
4. **It helps in analyzing the supplier performance and quality over time**. Big Data analytics can be used by the manufacturers are able to keep an eye on the product quality and the accuracy of delivery, which helps in suppliers receive the products on time. It also helps in maintaining quality [4].
5. **It helps in measuring the minute details of machinery and processes**. With the help of sensors, the information of each process is provided to the operations managers. Thus, it shows the quality and efficiency of each machine and its operators [4].
6. **It helps companies understand and sell the most profitable product configuration which has the least effect on the production** [4].
7. **It helps in managing the supply chain risk**. The supply chain is the key risk area where companies are quite concerned. Big Data analytics may help companies overcome the supply chain challenges. For example, the predictive analysis may help companies find out the probabilities of delay and hence companies may work accordingly [3].
8. **It helps in quantifying how production may influence the financial performance**. It has always been a problem for firms to set a connection between the daily production to the financial performance. But with the help of Big Data, the scenario is changing. With Big Data companies may have live information about a factory floor functioning and can help them scale the operations [5–12] better [4].
9. **It helps companies achieve the goal of Customer Satisfaction**. With the help of Big Data companies are able to keep a track of all the problems faced by their customers while using their products and thus, companies may work to find the solutions to the problems. It also helps companies analyze the benefits being provided by the other organizations and thus they can improve their customer services which finally leads to an increase in profit [4].

3 Surveys

The following survey has been done by McKinsey and Company illustrating how Big Data and advanced analytics are helping companies to streamline the production chains through finding the most important processes by measuring the process performance, and the thus help taking decisions to improve them continually [4] (Fig. 1).

By making use of Big Data analytics we can reduce the flaws during processing, increase the organization's efficiency and lead to efficient management of time and

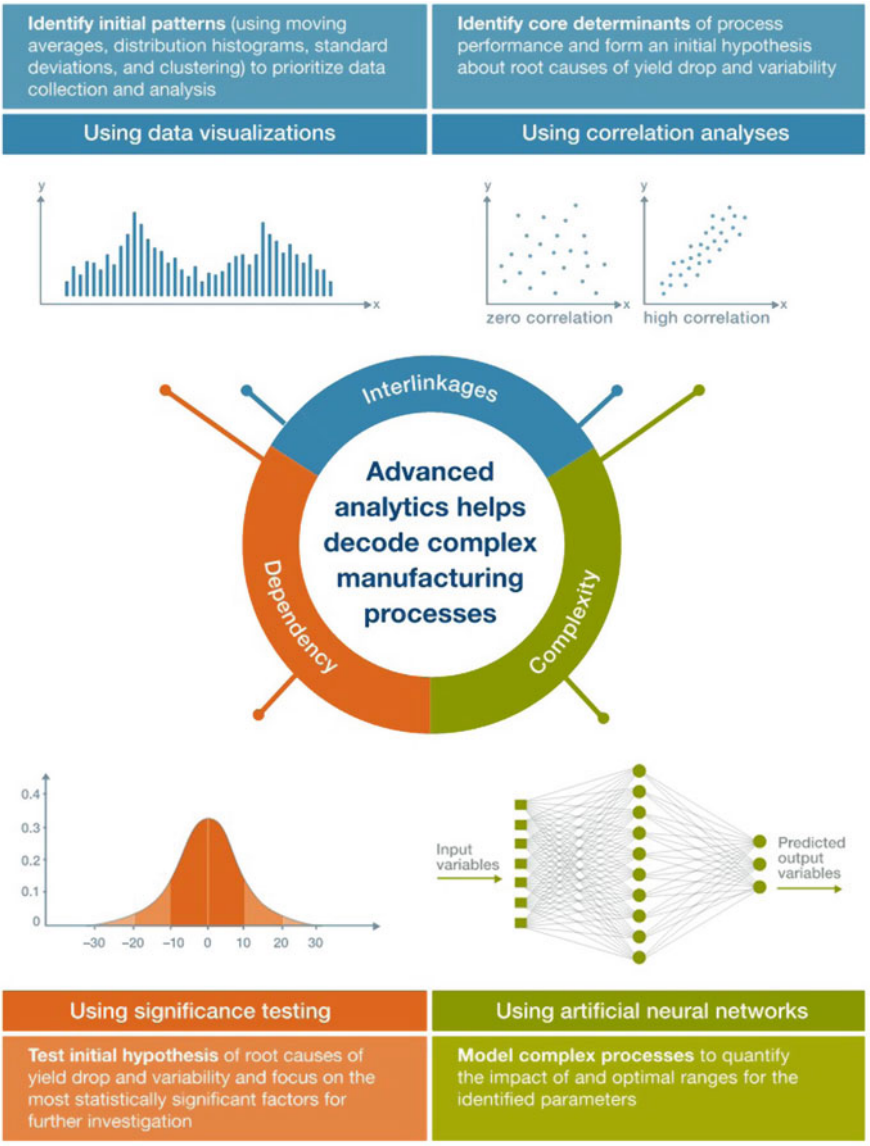


Fig. 1 Survey conducted by McKinsey [4]

money. **Tata Consultancy Services** conducted a survey in which it asked various organizations to rate the benefits of Big Data in manufacturing on a scale of one to five [13]:

- Quality of product and tracking defects—3.37
- Planning the supply mechanism—3.34

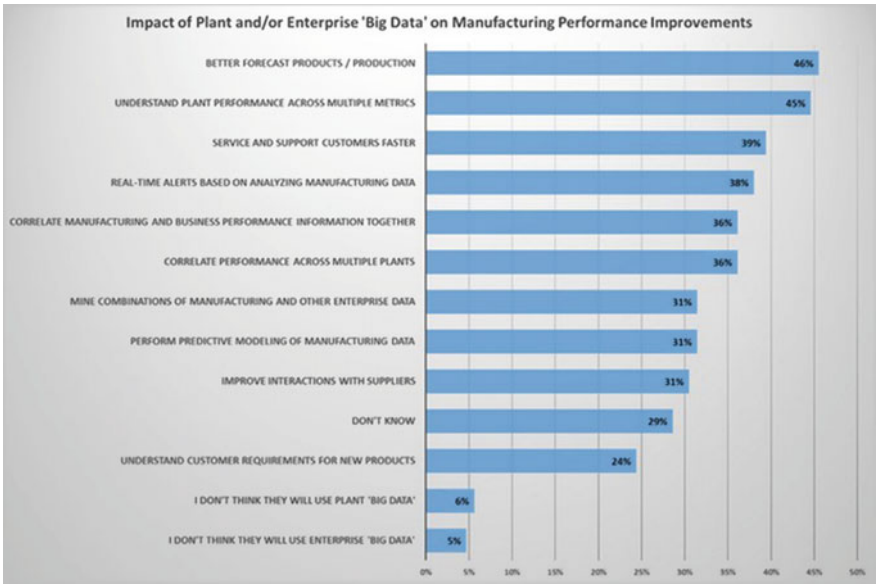


Fig. 2 Survey conducted by LNS research [4]

- Defect tracking of manufacturing processes—3.30
- Supplier, components and parts defect tracking—3.11
- Supply the performance data regarding the contract negotiations—3.08
- Output Forecast—3.03
- Increase in the energy efficiency—2.97
- Testing of products and simulation of new manufacturing processes—2.88
- To bring changes in manufacturing support—2.75 [14]

The following graph shows the findings of a recent survey conducted by **LNS Research and MESA International** so to analyze the processes where big data is showing the greatest improvements in the performance today [4] (Fig. 2).

4 Observations

According to the above-mentioned surveys and the studies I have done about Big Data and its applications, it shows a significant role Big Data can play in make in India. The results have been quite positive and they show how beneficial it is to implement Big Data. There has been a gradual increase in the production of the companies after implementing Big Data in their companies. Further factors like understanding plants performance across multiple matrices [15] have gone up significantly. Its one of the biggest impact is on customer services and support

which is one of the most important factor for companies. Customers focus a lot on the customer services and thus helps in building company's reputation. Many other factors have also shown a steady increase and there are many other factors that will show a gradual increase with time.

Also, manufacturers have given good ratings for the benefits of big data in various aspects of manufacturing. The benefits like product quality, defects ratings and supply planning, etc. have been given a good rating by various companies. Thus, it shows that manufacturers have understood the value of Big Data and are in favour of using it. The time is near when each and every company would be using the Big Data technology. Many global organizations have supported the use of Big Data.

Big Data and Make in India together can help in building a new India. This will benefit each and every company which invests in make in India. Big Data has worked in other countries and it will work in India too. It will help in understanding the need of people and the type of products needed to be built. Many world level surveys support Big Data in manufacturing. The best thing about Big data is that it is not for a specific industry. Thus, it can lead to overall industrial growth in the country with benefits for everyone ranging from customers to manufacturers. Big Data is a new concept and it will take time for companies to adapt to it.

But results show what Big Data can do to Manufacturing. The more industrialized a nation is, the more developed it is. Hence, Big Data and Make in India ds to a bright future of India. Time is near when we hope to see India in the top nations of the world in aspect.

5 Conclusion

The surveys mentioned above shows what Big Data can do in the field of manufacturing. It has lead to the growth in the production as well as lead to the increase in the customer satisfaction. Many MNCs are already using it and they have seen a steady growth in their production. Big Data also helps in managing large data set of employees. The amount of benefits Big Data provides, easily overcomes its one or two drawbacks. Make in India is one of the best initiatives of Indian government and it has lead to large investments in India and the amount of investments will surely go on increasing. India needs industrialization to increase jobs and to increase the growth rate. Big Data and Make in India combined together can bring an industrial revolution in India. So, if implemented properly Big Data will change the manufacturing industry in India and soon we will see India in top industrialized nations of the world. The scope of Big Data is immense in India and companies who are not making use of Make in India should learn from the companies who are already using Big Data. It will surely take time combining Make in India and Big Data, companies would have to invest more but the benefits would be great and long-term.

References

1. Make in India, Wikipedia (2016). https://en.wikipedia.org/wiki/Make_in_India. Accessed 2016 Apr 2
2. Big Data, Wikipedia, 3 April 2016. https://en.wikipedia.org/wiki/Big_data. Accessed 3 Apr 2016
3. Krishnan, A.: Applications of Big Data: manufacturing and governance. In: Digit Fast Track, A 9.9. Media Publication (2016)
4. Columbus, L.: Ten ways big data is revolutionizing manufacturing. Forbes, 28 November 2014. http://www.forbes.com/sites/louiscolumbus/2014/11/28/ten-ways-big-data-is-revolutionizing-manufacturing/#18309e3e7826. Accessed 1 Apr 2016
5. A Passion for Research (2015). https://softwarestrategiesblog.com/tag/cloud-computing/. Accessed 2 Apr 2016
6. Zikopoulos, P., Eaton, C.: Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, 1st edn. McGraw-Hill (2011)
7. Rajpathak, T., Narsingpurkar, A.: Managing knowledge from Big Data analytics in product development. Tata Consultancy Services
8. Hu, H., Wen, Y., Chua, T.S., Li, X.: Toward Scalable Systems for Big Data analytics: A Technology Tutorial. IEEE (2014)
9. McGuire, T., Manyika, J., Chui, M.: Why Big Data is the New Competitive Advantage (2012)
10. Joseph, J., Sharif, O., Kumar, A., Gadkari, S., Mohan, A.: Using Big Data for Machine Learning Analytics in Manufacturing. Tata Consultancy Services (2014)
11. Turner: Business Advantage Announces Results of Worldwide CAD Trends Survey. Business Advantage Company (2014)
12. MKGI: The Internet of Things. McKinsey Global Institute
13. Big Data Study: Tata Consultancy Services, 2010–2014. http://180.87.41.34/big-data-study/manufacturing-big-data-benefits-challenges/. Accessed 31 Mar 2016
14. 4 Big Data Use Cases in the Manufacturing Industry: Paragon Procurement, 16 9 2014. http://www.procurementprofessionals.org/4-big-data-use-cases-manufacturing-industry/. 1 Apr 2016
15. Wang, L., Alexander, C.A.: Big data in design and manufacturing engineering. Am. J. Eng. Appl. Sci. 11 (2015)

Agent-Based Wormhole Attack Detection and Prevention Algorithm in the Cloud Network Using MapReduce Technique

Priyanka Verma, Shashikala Tapaswi and W. Wilfred Godfrey

Abstract In day-to-day life, various cloud-based services are being used for disparate purposes because of its perpetuity and diverse dexterity. The cloud computing system has some affinity for distributed systems as both use the various features of networking. While using the cloud-based services, a colossal amount of data comes into picture, which upsurges the data traffic in the network. A cloud environment for this huge amount of data, it is challenging to secure it from various kinds of attacks. So an extensible and decisive threat monitoring, detection and prevention system is required for providing a highly securable infrastructure. The research have been done so far focused only monitoring the network for malicious activities, but not detecting and preventing any particular attack. This paper concentrates on a special attack called a wormhole attack on the cloud computing network. The paper proposes a scheme which detects and prevents the wormhole attack using agent nodes in the network. Multiple agent nodes having IDS capability are used to monitor the behavior of other nodes by analyzing the network traffic. Analysis of traffic is done from the following parameters: hop count, time delay, data packet's path. In cloud network, since the huge amount of traffic is captured by IDS, the work incorporates MapReduce technique to make the approach effectively. The Map procedure concurrently processes the data using the key value pair <k, v>, in our case key values are the parameters used for traffic analysis. The reduce procedure works on the basis of threshold defined for the parameters produces the list of nodes, which are creating wormhole attack in the network. The cornerstone of the proposed approach is, it does not impose any

P. Verma (✉) · S. Tapaswi · W. Wilfred Godfrey
Atal Bihari Vajpai—Indian Institute of Information Technology, Gwalior, MP, India
e-mail: 303priyanka.verma@gmail.com

S. Tapaswi
e-mail: stapaswi@iiitm.ac.in

W. Wilfred Godfrey
e-mail: godfrey@iiitm.ac.in

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_43

burden on the server, effective for all cloud models. For prevention of the wormhole attack, ID's of malevolent nodes is broadcasted to other nodes that the given nodes are detected as wormhole nodes.

Keywords Cloud computing · MapReduce · Wormhole attack · Intrusion detection system (IDS)

1 Introduction

Cloud computing is an unborn technology gaining its popularity because of its contrive services to the users for attaining or for using various applications. Diverse cloud-based services are used in day-to-day life. The cloud provides the facility for using the resources from a shared pool from anywhere at any time, provided that there must the availability of Internet. As a cloud is providing an ample number of facilities in a very effortless manner, so a colossal number of users are using many clouds-based services that in turn upsurge the data traffic in the network. The network had to deal with the heavy amount of data, and this causes various security challenges [1] including various types of attacks on the network.

The network is exposed to various kinds of attack. These attacks can be the active attack or passive attacks. Some of them are wormhole attacked, Sybil attack, black hole attack, DOS attack and much more. It is tough to detect a particular attack on the network, and a wormhole is one of the attacks which is very difficult to detect. As in cloud, there is an enormous amount of data which travels through the network, so for detection of any attack an efficient data analysis technique like MapReduce is required. Till now various research have been done in the field of only for monitoring the network for any malicious activity, but research is to be done to find a particular type of attack. The paper proposes a scheme, which detects and prevents the wormhole attack using agent nodes in the network with the help of MapReduce.

Wormhole attack [2, 3] is one of the most active and dangerous attacks, in MANET which can be easily executed. In this type of attack, a virtual link has been created that will divert all the traffic through that link and drop or relay all the packets in the network, and the packets will never reach its destination. In wormhole attack, two nodes called wormhole nodes or malicious nodes create the tunnel at separated distant points within MANET [4, 5]. The first wormhole node takes all the packets by showing that it is having the shortest path towards the destination, and afterward forwards all the incoming packets to the other wormhole node and this node will drop all the packets.

MapReduce is a programming model used by Hadoop [6] systems for the processing of massive data. MapReduce is based on two procedures Map procedure and the Reduce procedure. To speed up the process, the data are divided into various chunks and can be processed parallel. Map procedure is used to draw out the desired results from the data set, it uses a key/value pair. Based on these

key/value pair and the programming for the particular search, map procedure will list all the related data at each chunk. Reduce procedure is always executed after the Map procedure; it takes the output of map procedure and aggregates the result of each chunk and apply some threshold function to produce the final result.

The reminder section of this paper is organized as: Sect. 2 presents the literature survey of the work related to attacks in cloud network and its detection techniques. Section 3 presents the proposed work and its flow graph. Section 4 gives the conclusion of the paper.

2 Related Work

For any critical infrastructure, security is of primary concern. So an extensible and decisive threat monitoring, detection and prevention system is required. When the massive amount of data comes into the picture in these critical infrastructures, the problem of security is becoming more dangerous. So to come up with these challenges paper [7] proposes a system for network monitoring and threat detection based on cloud computing. The real-time data like network traffic and system logs are given as input to the MapReduce framework after data processing output results are stored in a MySQL database. If any malicious activity is detected, then an alarm will be rung to inform the administrator, and monitoring system will generate a pop-up, a warning window to address the threat.

For solving scalability issues federating multiple clouds had come out to be a good solution, but with this, it had to deal with possible vulnerabilities. In paper [8] an agent-based system in which agents overseers the publishes/subscribe services is proposed. The agents used are trained to act as the investigator in order to gather information about the quality of service parameters and performance metrics in the middleware infrastructure. Agents can communicate for exchanging data. To detect and reject the expected attacks in cloud federated system they trigger alerts or other actions. In future, they are planning to work on the detection of any particular type of DOS attack.

For preventing the network dimout and maintaining the availability of resources, anomaly detection is essential. But with the increased growth of the Internet traffic, it is very difficult to identify the harmful traffic in real time. In paper [9] scheme uses benefits of distributed computing approaches for real-time analysis of non-sampled internet traffic is proposed and also uses Hadoop for network anomaly detection. The paper proposes a Hashdoop, a MapReduce framework that splits the traffic with a hash function. Using hash function traffic is divided into splits and at each split, detector identifies the anomalies; thereafter anomalies are then collected and reported to the network operator.

In MANET due to the lack of a central point of control, networks are more vulnerable to attack. A wormhole is one of the most severe attacks which is very difficult to identify. In the paper [9] they had analyzed the existing techniques and their limitations for identifying the attack. They had discussed various MANET

features like location, time, hop count, neighborhood, data packets, route reply, and route request for the detection of wormhole attack. They had also discussed the pros and cons of each feature in detail and also provide the possible limitations of the IDS system. The work provides the basis to build efficient IDS to detect wormhole attack.

As cloud computing is gaining its popularity because of its various services, so is used by many organizations. With this lot of security incidents and new kinds of vulnerabilities also come into the picture. The paper [10] gives an overview of the possible cloud computing attacks and provides a solution to these attacks to some extent. But still so many attacks are uncovered and need to be detected by some efficient manner to secure the cloud network.

As many papers are using various techniques for attack detection and prevention, but none of them are providing the specific attack detection and prevention technique. But many of them had considered detecting specific attacks in their future work. Gaining the inspiration, the paper proposes a detection and prevention technique for specific kind of attack called a wormhole attack on the network between the cloud server and the user. In the paper proposed scheme is using agent node [11] with IDS [12] capability and the MapReduce framework for the detection of wormhole nodes.

3 Proposed Work

3.1 Overview

The paper proposes a scheme in which wormhole attack detection and prevention between cloud servers and the cloud user's system is presented. As with growing use of cloud technology to a huge number of requests from the user are given to cloud servers which server has to process. This creates massive amounts of data, which travel through the network. So there is need to secure the network from various kinds of attack. In this paper, we are proposing a scheme in which Monitoring agent with the IDS capability is used in the network, which monitors all the network traffic and detects the nodes performing any malicious activities in the database (Fig. 1).

The proposed work has used the MapReduce feature as well for fast and efficient detection. The malicious nodes stored in the database are given to the Map phase of the system. The Map will use the detection algorithm and list the expected wormhole nodes. The Reduce phase will recollect the expected wormhole nodes from each chunk and prepare a list and apply a threshold function to identify the actual wormhole nodes. Now the node ID's performing wormhole attack is broadcasted by the monitoring agent to all the nodes in the network. So the nodes using the path which is having wormhole nodes in between will change the path and chooses the alternative path for data transfer which does not contain the wormhole nodes.

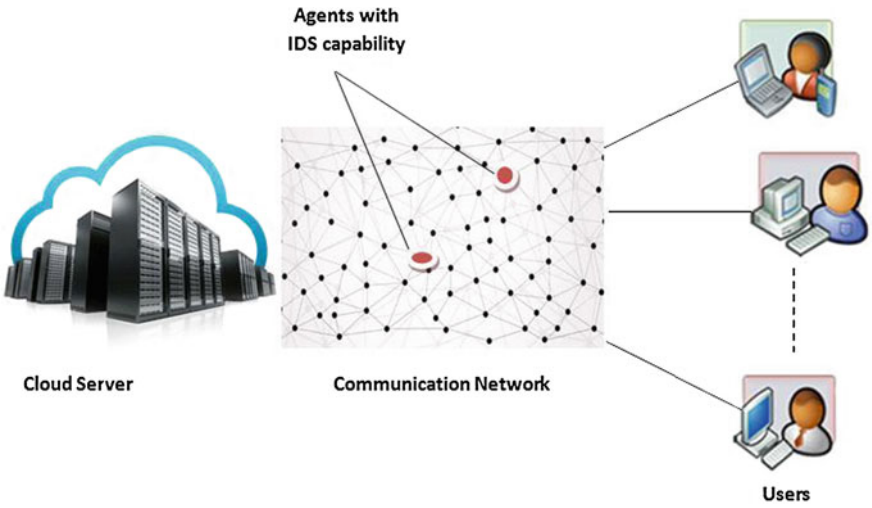


Fig. 1 Network with agents having IDS capability

3.2 Flowchart of Proposed Work

See Fig. 2.

3.3 Workflow

Monitoring Agents: Monitoring agents are used in the network to monitor all the data traffic in the network. This paper proposes a scheme in which monitoring agent with IDS capability is used. To secure the network from the attacks, a large volume of data with their associated information like traveling path, hop count and the time delay is analyzed for the purpose of threat detection. The IDS uses these parameters for the selection of malicious nodes. The agent with this capability can detect the nodes performing some malicious activity and stores the node ID and its associated parameter in the database for further detection of wormhole nodes out of these listed malicious nodes. When Wormhole nodes are detected, after the MapReduce procedure these agents will broadcast the ID's of wormhole nodes in the network for prevention.

Map Procedure: A database of malicious nodes produced by the agent with the help of IDS is used as input by the Map function. First, the Map divides this database into "N" chunks for the fast and parallel processing, then for each chunk, we use the Function Map described below. In this for every pair of nodes, it checks the weather data is continuously captured by them or not. If the data is continuously captured by them and none of them is received, then we declare those pairs of nodes

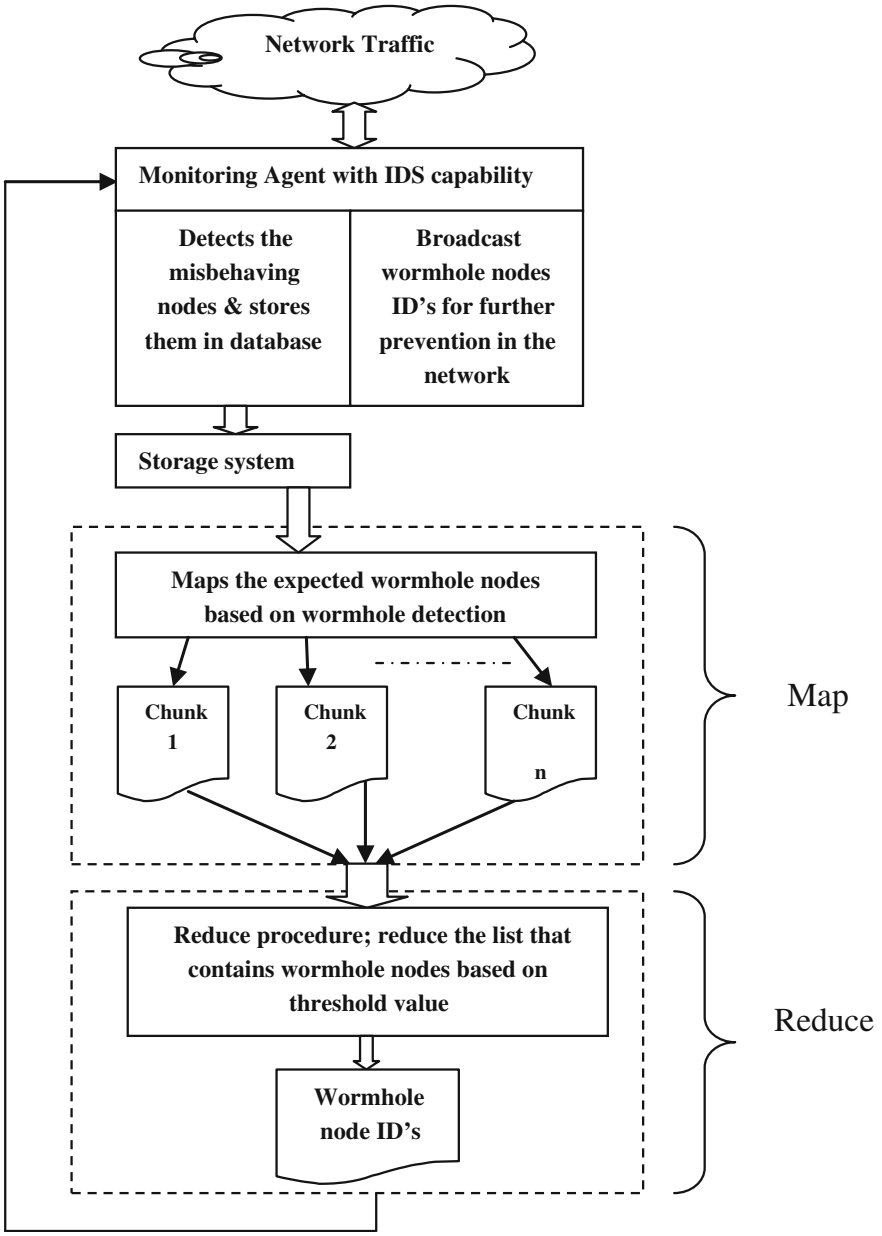


Fig. 2 Flowchart diagram of wormhole attack detection

as the suspected wormhole nodes. In every chunk, this procedure is applied and in turn, it enlists the suspected nodes in that chunk. Following Map function is used for the detection of suspicious wormhole nodes:

Function Map is

Input: List of all malicious nodes detected by
IDS Check for every pair of suspicious nodes
do
If (Sender ==Xi&& next hop ==Yi)
Check Yi forward data or not
If (Yi not forward any data & Yi! = receiver)
If (Xi and Yi continuously captures data)
Produce output {Xi and Yi}
Repeat
End function

In the above algorithm, the list of nodes are given as input, and for every pair of nodes say Xi and Yi it is checked that if Xi node is sending data and Yi is not forwarding the data further, then we check for whether the Yi is received or not. If the Yi is not the receiver and is also capturing the packet and not forwarding further to the intended receiver, then function gives out that these two nodes are suspected to be wormhole nodes. The procedure continues for every pair of nodes.

Reduce Procedure: In the reduce part the function, reduce will collaborate all the suspected wormhole nodes from each chunk as produced by Map procedure. After preparing the list it will apply a threshold function to all the pairs of nodes. The threshold function is defined for the maximum number of data captured between any pair of intermediate nodes. If any pair of node from the list produced by Reduce function, exceeds this threshold value, then that pair of node is detected as wormhole node. Finally, reduce will produce a final list of nodes detected as wormhole nodes after applying the threshold function. Below is the function Reduce used:

Function Reduce is

Input: List of expected pair of wormhole nodes (Xi & Yi)
For each input node (Xi & Yi)
do
Amount of data captured by (Xi & Yi) = Z
If (Z > Threshold value)
Produce output (Xi and Yi are wormhole nodes)
Repeat
End function

In the above algorithm, the list of Reduced expected pair of nodes is given as input and for each pair, it will check the amount of data captured between them. If this amount of data exceeds the threshold limit the pair of nodes is detected as wormhole node. The threshold function is specified as the maximum amount of data captured between any pair of intermediate nodes.

4 Conclusion

This paper proposed a scheme for the detection and prevention of wormhole attacks in the cloud network between cloud servers and the users. The proposed system had used the agent monitoring system in the network with IDS capability, the system also makes use of MapReduce technique for fast and efficient detection of the wormhole attack. The main advantage of the proposed work is that it does not imply any extra burden on the server because it is applied in the network in between the cloud servers and the users. The proposed work can be applied for all cloud models. In future research work can be extended to find other, specific attacks in the cloud infrastructure.

References

1. Ali, M., Khan, S.U., Vasilakos, A.V.: Security issues in cloud computing: opportunities and challenges. Inf. Sci. Elsevier **305**, 1 (2015)
2. Mahajan, V., Sethi, A.: Analysis of intrusion wormhole attack in MANET. In: Military Conference (2008)
3. Ahuja, R., Banga, A., Ahuja, P.: Performance evaluation and comparison of AODV and DSR routing protocols in MANETs under wormhole attack. In: IEEE Second International Conference on Image Information Processing ICIIP (2013)
4. Khalil, I., Bagchi, S., Shroff, N.B.: MOBIWORP: mitigation of the wormhole attack in mobile multihop wireless networks. In: Securecomm and Workshop IEEE (2006)
5. Jigalur, R.S., Bhushan, C.: Designing a secure architecture against wormhole attacks in wireless sensor networks. In: Computing, Communications and Networking Technologies (ICCCNT), IEEE (2013)
6. Lee, Y., Lee, Y.: Detecting DDoS attacks with Hadoop. In: ACM CoNEXT Student Workshop, Dec 2011
7. Chen, Z., Xu, G., Mahalingam, V., Ge, L., Nguyen, J., Yu, W., Lu, C.: A cloud computing based network monitoring and threat detection system for critical infrastructures. Big Data Res. Elsevier (2016)
8. Verma, P., Dhariwal, S., Tiwari, H.: Wormhole attack intrusion detection and prevention security scheme in MANET. Int. J. Comput. Appl. **105**(10), 0975–8887 (2014)
9. Fontugne, R., Mazel, J., Fukuda, K.: Hashdoop: a MapReduce framework for network anomaly detection. In: IEEE Conference Computer Communications Workshops (INFOCOM Workshop) (2014)
10. Singh, S., Pandey, B.K., Srivastava, R., Rawat, N., Rawat, P., Awantika: cloud computing attacks: a discussion with solutions. Open J. Mobile Cloud Comput. **1**(1) (2014)
11. Ficco, M., Tasquier, L., Aversa, R.: Agent-based intrusion detection for federated clouds. In: International Conference on Intelligent Networking and Collaborative Systems. IEEE (2014)
12. Aljarah, I., Ludwig, S.A.: Towards a scalable intrusion detection system based on parallel PSO clustering using MapReduce. In: GECCO Genetic and Evolutionary Computation Conference. ACM (2013)

Tourism Recommendation Using Machine Learning Approach

Anjali Dewangan and Rajdeep Chatterjee

Abstract Puri tourism has always remained as the best tourist spot in Odisha. Researchers and town planners have always taken steps in finding out for proper tourism recommendation. But always they have preferred the method of machine learning approach for the tour recommendation models. Some methods give good simulation data but sometimes artificial neural network (ANN) and regression analysis techniques give better results. In this paper, Puri tourism recommendation method has been modelled based on the SOM architect, and by revenue management system. Here, a complete comparison has been described between supervised and unsupervised machine learning technique for tourism recommendation in Puri.

Keywords ANN · SOM architect · Regression · Simulink block diagram Revenue management system

1 Introduction

Tourism is just like an industry for a country for not only its development from the cultural point of view but also adds good exchange program for development in terms of trade, industry and many more [1]. So, tourism forecasting is completely necessary for industry's contribution to the economical development of that region. So, it is very helpful for managers and government. Actually, government organizations use supervised machine learning techniques like regression analysis for achieving marketing targets and helps them to attain marketing potential stability. Managers use these techniques for determination of staffs and capacity for the study of financial projects to build new hotels and do town planning for tourism recommendation in a country. Taking these techniques into account e Tourism helps in

A. Dewangan (✉) · R. Chatterjee
School of Computer Engineering, KIIT University, Bhubaneswar, India
e-mail: er.anjalidewangan@gmail.com

R. Chatterjee
e-mail: cse.rajdeep@gmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_44

providing information service features such as travel agents, hotel and tourist spots. The main thing is how to optimize time, money and basic cost for food. Tourist needs more demands in short interval of time and never take interest for searching for too long time for an online assistant. So, to cross across this situation ANN and data mining are now being used widely. The essence of this paper is to symbolize a model of tourism based upon ANN and time series.

2 Models

2.1 Time Series Model

This has been extensively preferred in the regression analysis for the calculation of trend and seasonality analysis. This helps in the prediction of future tourist to a specific tourist spot. In this paper, this has been used for tourism analysis for Puri destination [2]. Here, in this method basically autoregressive integrated moving average technique (ARIMA) has been used to model a forecasting method for evaluation of fitness function along with that to choose the best error method for to measure the performance.

2.2 Artificial Neural Network (ANN) Model

The two methods which are widely used are supervised and unsupervised machine learning approaches. Neural network is widely used in forecasting and prediction of future response for the tourism industry. One of this parts used for analysis is multi-layer perceptron method. It is just like a bridge for input and output layers that is based upon the initial simple perceptron method with many branches of hidden neurons which helps in identifying the capacity for to learn MLP network [3]. In this method, past 10 years of tourist numbers have been collected from various agents and companies and from travel agents also. Then they are being simulated in SOM architecture methodologies.

3 ARIMA Method

It includes basically fitness function, genetic algorithm method which helps in regression analysis. Below we have got the fitting function for each trend and monthly variation of tourist coming to Odisha (Puri) [4]. Here basically three errors have been analyzed to fit the error terms in the regression analysis [5, 6]. Among these, three errors only have been selected based upon the performance for nearness

Table 1 Tourist in terms of thousand to Puri (Odisha)

2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	
112	115	145	171	196	204	242	284	315	340	360	417	January
118	126	150	180	196	188	233	277	301	318	342	391	February
132	141	178	193	236	235	267	317	356	362	406	419	March
129	135	163	181	235	227	269	313	348	396	396	461	April
121	125	172	183	229	234	270	318	355	363	420	472	May
135	149	178	218	243	264	315	374	422	435	472	535	June
148	170	199	230	264	302	364	413	465	491	548	622	July
148	170	199	242	272	239	347	405	467	505	559	606	August
136	158	184	209	237	259	312	355	404	404	463	508	September
119	133	162	191	211	229	274	306	347	359	407	461	October
104	114	146	172	180	203	237	271	305	310	362	390	November
118	140	166	194	201	229	278	306	336	337	405	432	December

to zero and these errors have been calculated for first ten data of tourist and then forecasted and fitted to regression accordingly.

$$MSE = \frac{1}{N}\sum_{r=1}^{N}\left(F_t - \overline{F_t}\right)^2. \tag{1}$$

$$MAE = \frac{1}{N}\sum_{r=1}^{N}\left(F_t - \overline{F_t}\right)^2. \tag{2}$$

$$MAPE = \frac{1}{N}\sum_{r=1}^{N}\left(\frac{F_t - \overline{F_t}}{F_t}\right). \tag{3}$$

where N is the size tourist time series data as per their arrival, F_t is the actual value and $\overline{F_t}$ is forecast value at time t for the year span between 2000 and 2011. MAPE is employed for the comparative study of the performance of forecast model, whereas MSE is represented as forecast analysis in SARIMA model and exponential model. Here the trend may be linear or quadratic depending upon seasonal component. In the year 2005, the monthly variation was in between 4.5 and 5 but after 1 month it rose to between 6 and 6.5 [7–9]. Below is the data for tourist used in our simulation from the past 10 years coming to Odisha (Puri).

Here in Table 1, the past 10 years tourist data has been taken into consideration which says about their numbers of Odisha tourism. These data are in terms of thousands. In these data, it has been revealed that in the year 2000 the number has been increased to a peak level and then decreased to a certain in the months from January to December. And it has been repeated in the same manner, so the trend has been fitted and then found out to be in a quadratic with linear fashion. Then its error analysis has been calculated and then fitted with response function to give the desired trend which helped to analyze the year forecasting in tourism for Puri (Table 2).

We did the calculation for first 10 years data of passengers who came to Puri and error analysis upon calculation do not show better calculation because though the

Table 2 Error calculation

MAE	MAPE	MSE
2.4676	0.022032	543.6652
2.4076	0.020403	491.232
2.2676	0.017179	389.5462
2.2976	0.017811	409.2222
2.3776	0.01965	467.1886
2.2376	0.016575	370.8781
2.1076	0.014241	300.1336
2.1076	0.014241	300.1336
2.2276	0.016379	364.8678
2.3976	0.020403	483.066

error in case of MAE is good than MSE but both are not having less error value so that our theoretical regression does not overlap with analytical data. It shows that MAPE has got good regression analysis in calculation [10–12].

4 Numerical Computation of Tourist Forecasting and Its Regression Analysis

In this paper, we have analyzed Mackey–Glass time-delay differential equation. Here in the equation with our tourist demand for tourism to Puri is like a non-periodic and non-convergent time series that is sensitive to initial conditions. As per the equation:

$$\frac{dx(t)}{dt} = \frac{0.2x(t-tau)}{\left(1+x(t+tau)^{10}\right)} - 0.1x(t), \quad (\mathrm{x(t)} = 0, \text{ where } \mathrm{t} < 0) \tag{4}$$

Here, tau is the time delay for our forecasting of tourist number [13]. This x denotes the *i*th row and *k*th column in the regression matrix for tourist number. Then it is fitted with regression model which have an explicit forecasting mechanism and well-defined stationary, invertibility requirements. Then it takes the form of linear form as used in statistics where response form for regression fitting has been got. The equation in fitting takes the form as follows: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_3 x_2^2 + \in$. Here β is the regression coefficient and ε is the MAPE error, and x is the regression matrix and y is the response. Here in the figures in the error analysis, y-axis represents the error and the x-axis represents the number of passengers in terms of thousands [14–16]. Among the entire error fitting, MAPE has been decayed rapidly giving less error in comparison to others. Here for error analysis first 10 number of passenger samples have been taken into account.

5 SOM Architecture Methodologies

Self Organizing Maps (SOM) mainly comprises of three things; first one is input layer, middle one is a competitive layer and the last one is output layer. Here neural network has been used for simulation. For training purpose trainer is applied. Input vectors have been classified and they are been grouped in the input space layer. Input layer is totally different from middle competitive layer because this middle layer helps in recognizing neighbouring sections from the input layers. So, both distribution and topology of input vectors are done by this methodology.

Hereby the help of topology functions, the neurons are arranged in physical positions. The topology functions are gridtop, hextop or randtop that helps in arranging the neurons in a grid, random or hexagonal way. Distance function is first

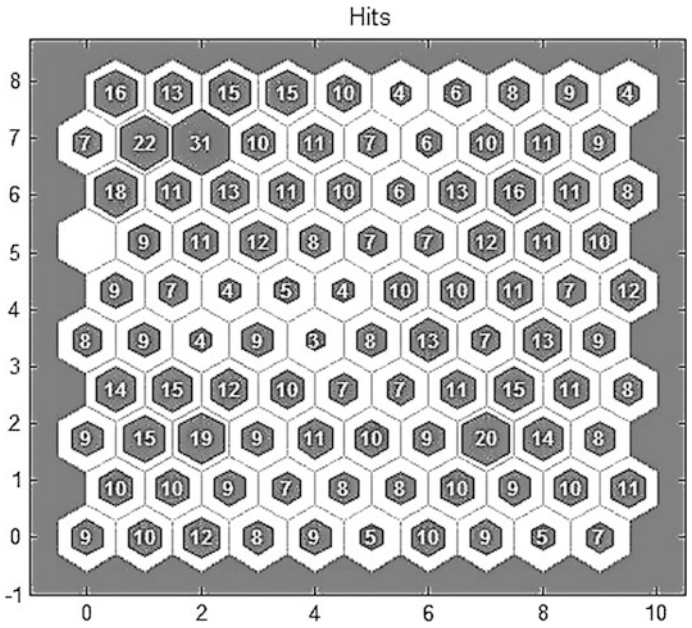


Fig. 1 SOM sample hits

of all calculated between neurons and then they are positioned accordingly. Here weight function has been taken into consideration for each input vector. With respect to arrival rate, these vectors denote service rate and waiting time of tourist. This design is a kind of competitive network, except bias being utilized. The competitive copy function creates a selection corresponding to i*, the winning neuron for output aspect $a1_i$. Other outcomes of all elements in a1 are 0. Neurons near the winning neuron are up-to-date along with the receiving neuron. Here, the competition of neuron in winning is analogous to tourist coming to their choice hotels and neurons have been made similar to the waiting time of each tourist with their preference budget.

In Fig. 1, it tells about the SOM sample hits which consist of large and small weights [17, 18]. The larger weights are in the blue colour. These connect the input to their respective neuron by weights. In the event, the connection patterns of two inputs are incredibly similar, you can assume that the inputs were highly related. In this instance, input 1 has connections that are incredibly different than the ones from input 2.

If perhaps the input space is high dimensional, you are not able to visualize all the weight loads at the same time. Here, the black colourings with hexagons represent the neurons. The red lines have been linked to neighbouring neurons. The

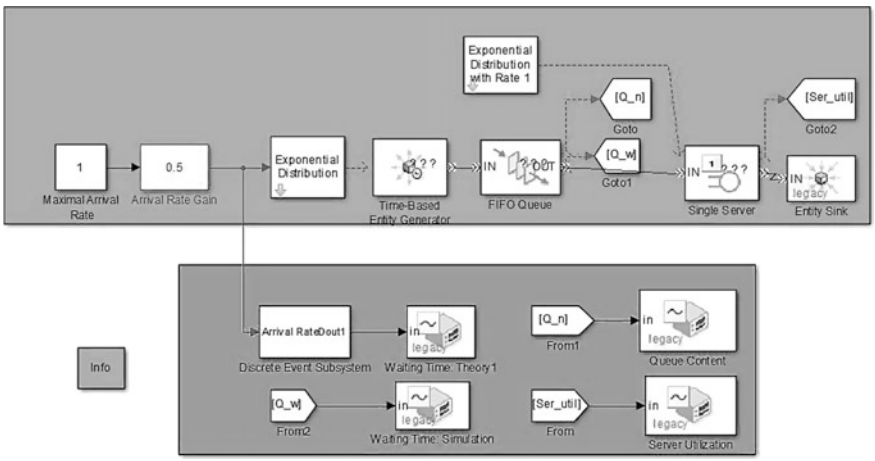


Fig. 2 Simulink block diagram for hotel agent in tourism field

Table 3 Hotel package in Puri

Hotel package (#1)	Hotel package (#2)	Hotel package (#3)	Hotel package (#4)	Hotel package (#5)	Hotel package (#6)
Season: summer	Season: summer	Season: summer	Season: winter	Season: winter	Season: winter
Destination: India, Puri	Destination: India, Puri	Destination: India, Puri	Destination: India, Puri	Destination: India, Puri	Destination: India, Puri
Hotel Stars: 5 (Mayfair Waves)	Hotel Stars: 4 (Fort Mahodadhi)	Hotel Stars: 3 (Shakti International)	Hotel Stars: 5 (Mayfair Heritage)	Hotel Stars: 4 (Hans Coco Palms)	Hotel Stars: 3 (Swargadwara Hotel)
Number of days: 7	Number of days: 7	Number of days: 7	Number of days: 7	Number of days: 7	Number of days: 7
Min price: 25000	Min price: 19000	Min price: 10000	Min price: 22000	Min price: 18500	Min price: 15000

colourings in the regions made up of the red lines show the distances between neurons. The dark colours symbolize large distances and light colours represent small ranges. Here, the distances reveal the tourist choosing their hotel as per their budget in tourism recommendation model. Here clustering of data has been split into two parts. Here the smaller weights are interconnected with larger weights; the darker bands represent larger weights. In Table 3, we have taken the hotels situated at Puri for tourism recommendation model as per tourist staying and lodging choice with their respective cost.

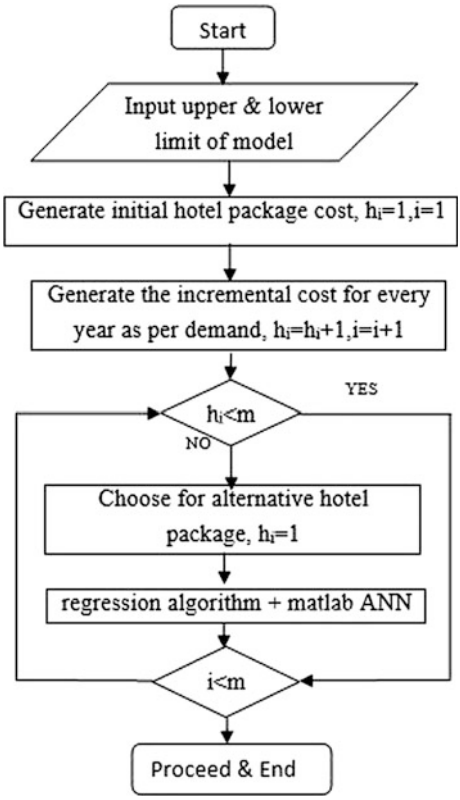


Fig. 3 Flowchart of tourism recommendation model

6 Tourism Recommended Model

The room allocation has been made for a hotel based on the economic principles. The key monetary principles are put on pricing, executes to functions like optimization and forecasting and controlling rooms inventory. In this paper, we have divided into 3 parts, namely, travel agent, hotel booking agent and governmental bodies and they are reservation system, DB, and our revenue management system. They all are linked with hotel reservation agent and revenue manager. Figure 2 is for tourism recommendation model which reveals the entire tourism for puri destination associated with passengers visiting puri which has been simulated through MATLAB Simulink taking first come and first serve technique into consideration.

Here, the important things are the maximum arrival rate and arrival gain rate which has been linked to the exponential distribution of tourist or passengers coming throughout the year from 2001 to 2011. Then FIFO (First Input and First Output) has been taken into consideration [19]. Here queue content and server utilization have been linked with waiting time with exponential distribution associated with tourist. This exponential function has been developed with passengers in between single month along each row for different years. For example, if we take

the case of month January, then tourist number has been increased at an exponential rate from 112 to 417. The flowchart for the tourism recommendation model is depicted in Fig. 3.

Steps used in Algorithm

Input: 2 parameters (queue content + server utilization)

1. Initialize maximum arrival rate and arrival gain rate (0.5, 0.6 and 0.7).
2. For each gain rate fit that to exponential distribution and pass it to FIFO queue by time-based entity generator.
3. Set FIFO.
4. Generate cost for the hotel; $h_i = h_i + 1$, if $h_i < m$, go to (queue content + server utilization).
5. Increment $i = i + 1$.
6. Else choose for alternative hotel package, $h_i = 1$.
7. Compute x which represents the tourist number.
8. Update y as response and add the smallest possible error (MAPE)
9. End for.
10. Output: set of increment numbers corresponding to tourism growth rate.

7 Case Study

The above algorithm has been analyzed with Figs. 2 and 4. Here 'h_i' and 'm' represent the cost of hotel package and tourist passenger demand cost. If the cost of hotel package is less than tourist demand, then it will work out or else the tourist has to look for an alternative solution [20]. Here the model after the simulation has been observed with the analytical values. The arrival gain rate for queuing model has been simulated with the range interval of 0.1–0.99. So here we have taken the arrival gain rate as 0.5, 0.6 and 0.7 for tourist when they come to the hotel for booking. Accordingly, we have done forecasting.

In Fig. 5, the first one has been simulated with arrival gain rate of 0.5 and the waiting time has been at first rose to 1.5 and then came to steady-state waiting for time 1 [21, 22]. Similarly in the second figure when simulated to arrival gain rate of 0.6 then waiting time rose to 2.5 and then decreased to 1.2 and then it came to a steady state of 1.5. And when in the third figure we did the simulation taking arrival rate gain of 0.7 waiting time rose to 5. Correspondingly, we have the server utilization keeps on increasing with an increment of waiting time. In the first figure of server utilization was first increased to 0.57 and then it comes to steady-state value of 0.5. Similarly, for second one it again increased to 0.67 and then it came to

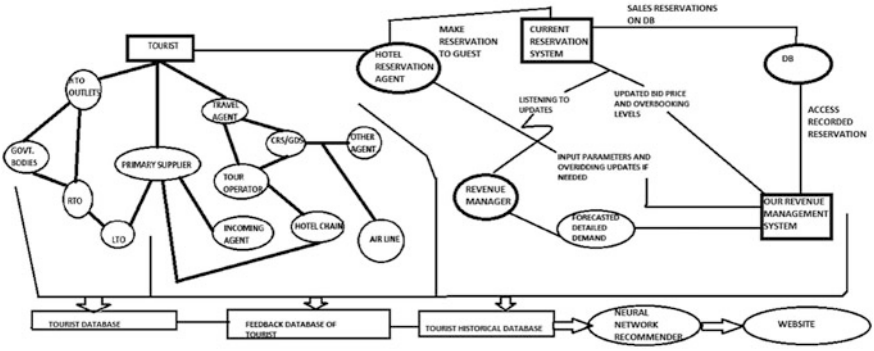
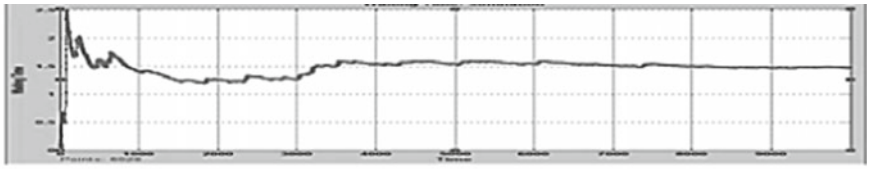
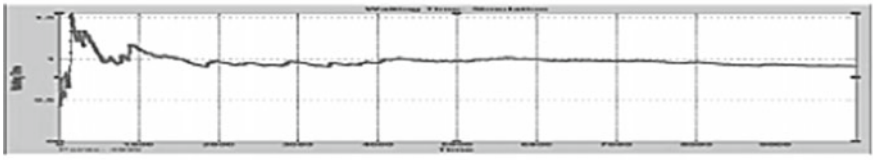


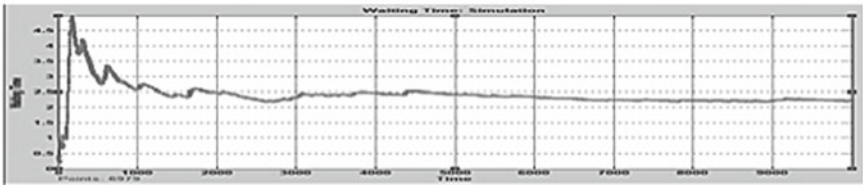
Fig. 4 Layout for hotel agent, travel agent, tourist and governmental agencies



(i). Waiting Time vs. Time



(ii). Waiting Time vs. Time.



(iii). Waiting Time vs. Time

Fig. 5 **a** Waiting time versus time. **b** Waiting time versus time. **c** Waiting time versus time

steady state of 0.6. And hence similarly for arrival gain rate for 0.7, again the server utilization gets increased to 0.75 and then it came to steady state of 0.7. So, the agents play a very important role in the economic development of that region. The more the tourist the more will be the budget and hence harmony between the native's places of tourist and destination for travelling spot area will take to a friendship bond for which economic growth rate will develop. The values obtained from simulation model are shown in Table 4.

Table 4 Values obtained from simulation model

Waiting time	Waiting time	Waiting time
1	0.5	0.5
1.5	0.6	0.6
2.5	0.7	0.7

8 Conclusion

We have compared the above simulation taking artificial neural network and regression technique into account. But among all methods the algorithm based on neural network and self-organized map depicts the time-saving method with lots of data. And on the other side regression has helped in forecasting the tourist incoming rate in an exponential and linear manner.

References

1. Zhang, G.P.: Neural networks in business forecasting, idea group inc.. In: Law, R., Pine, R. (eds.) Tourism Demand Forecasting for the Tourism Industry: A Neural Network Approach, Ch. 6 (2004)
2. Lim, C., McAleer, M.: Time series forecasts of inter-national travel demand for Australia. Tour. Manage. **23**, 389–396 (2002)
3. Goh, C., Law, R.: Modeling and forecasting tourism demand for arrivals with stochastic nonstationary seasonality and intervention. Tour. Manage. **23**, 499–510 (2002)
4. Law, R., Au, N.: Back-propagation learning in improving the accuracy of neural network-based tourism demand forecasting. Tour. Manage. **21**, 331–340 (2000)
5. Elmaghraby, W., Keskinocak, P.: Dynamic pricing in the presence of inventory considerations: research overview, current practices, and future directions. Manage. Sci. **49**(10), 1287–1309 (2003)
6. Canina, L., Carvell, S.: Lodging demand for urban hotels in major metropolitan markets. J. Hospitality Tour. Res. **29**(3), 291–311 (2005)
7. Gillen, D.W., Morrison, W.G., Stewart, C.: Air travel demand elasticities—concepts, issues and measurement. Technical Report, Department of Finance Canada (2004)
8. Cross, R.G., Higbie, J.A., Cross, Z.N.: Milestones in the application of analytical pricing and revenue management. J. Rev. Pricing Manage. https://doi.org/10.1057/rpm.2010.39 (2010)
9. Scholkopf, B., Smola, A.J.: Learning with Kernels—Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press (2001)
10. Smola, A.J., Scholkopf, B.: A Tutorial on Support Vector Regression NeuroCOLT. Technical Report, TR-98-030 (2003)
11. Canu, S., Grandvalet, Y., Guigue, V., Rakotomamonjy, A.: SVM and Kernel Methods Matlab Toolbox, Perception Systems et Information. INSA de Rouen, Rouen, France (2005)
12. Li, G., Song, H., Witt, S.F.: Time-varying parameter and fixed parameter linear AIDS—an application to tourism demand forecasting. Int. J. Forecast. **22**(1), 57–71 (2006)
13. Lim, C., Mcaleer, M.: Asian tourism to Australia. Ann. Tour. Res. **28**(1), 68–82 (2001)
14. Rasmussen, C.E., Williams, C.K.L.: Gaussian Processes for Machine Learning. MIT Press (2006)
15. U.S. Lodging Industry Results, tech. rep., Smith Travel Research (2007)
16. Monthly traffic analysis, Technical Report, International Air Transport Association (2008)

17. Rushmore, S.: Mid-rate extended-stay provides best return. Hotels **34**(5), 42 (2000)
18. Baker, T.K., Collier, D.A.: The benefits of optimizing prices to manage demand in hotel revenue management systems. Prod. Oper. Manage. **12**(4), 502–518 (2003)
19. Bellman, R.E.: Dynamic Programming, p. 862270. Dover Publications, Incorporated (2003)
20. Schwarz, Z.: Changes in hotel guests willingness to pay as the date of stay draws closer. J. Hospitality Tour. Res. **24**(2), 180–198 (2000)
21. Jeffrey, D., Barden, R.R.D.: Monitoring hotel performance using occupancy time-series analysis—the concept of occupancy time-series analysis—the concept of occupancy performance space. Int. J. Tour. Res. **2**(6), 383–402 (2000)
22. Benghalia, M., Wang, P.P.: Intelligent system to support judgmental business forecasting—the case of estimating hotel room demand. IEEE Trans. Fuzzy Syst. **8**(4), 380–397 (2000)

A Secure Clustering Technique for Unstructured and Uncertain Big Data

Md Tabrez Nafis and Ranjit Biswas

Abstract In clustering process, a set of patterns is separated into disjoint and identical significant groups. Faster data analyzing is one of the important aspects of clustering method. A number of works have been reported by various authors to optimize its multidimensionality toward distinct big data sets. However, the existing techniques are unlabeled to offer an optimal solution with regards to bunching high dimensional information set as their multifaceted nature tends to make things more hazardous while quantities of measurements are included. Especially, for uncertain and unstructured data it produces very poor results. Apart from that, the searching of data from the cloud or storage systems and data security are also important aspects. The current techniques also fail to offer proper care or solution in these matters. In this paper, the authors propose a secure clustering technique for unstructured and uncertain big data, and design an algorithm SCTA. This proposed technique does also offer high dimensionally for distinct types of big data. It includes SDES encryption technique for securing data as well as to maintain low complexity. Consequently, it includes a data searching algorithm from the cloud or storage systems to search data efficiently with low complexity.

Keywords Cluster · Unstructured data · Uncertain data · SDES

1 Introduction

According to [7], in this era of big data, uses of distinct big data files in distinct real-life applications are increasing very rapidly with the massive enhancement of high-speed Internet availability. Substantial amounts of data are generated every

M. T. Nafis (✉) · R. Biswas
Department of Computer Science & Engineering,
Jamia Hamdard University, New Delhi, India
e-mail: tabrez.nafis@gmail.com

R. Biswas
e-mail: ranjitbiswas@yahoo.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_45

day and transferred over the Internet by distinct social sites like Facebook, Twitter, search engine like Google, Yahoo, Amazon, knowledge sharing sites like Wikipedia, distinct research organizations in the form of distinct scientific or survey information, banking as well as financial organization in the form of various transactional as well as customer information, governmental surveillance data, and so on [4]. There are various challenges that exist during management, storage as well as the transfer of a huge volume of raw data in the form of structured, unstructured or semi-structured data. Data losses, issues of data privacy and integrity are the main challenges in managing such data. Data loss can occur due to improper space management, system error, unethical interferences of illicit third parties, inefficient data searching from the cloud or storage system and so on. Similarly, data privacy and integrity may suffer security attack, transportation errors or inefficient data management system [2].

Hence to manage as well as to store such huge amount of structured, unstructured, or semi-structured data in the remote clouds or remote storage systems require special care. A number of efforts have been made by the researchers to manage, transfer, and store such big data file efficiently and securely. As the consequence of it, diverse clustering techniques have been introduced to optimize the management and security of large-scale data [1]. However, the existing popular clustering algorithms are inherently difficult to parallelize, and also they are a poor performer in computation at large-scale multidimensional datasets [6].

In this paper, our work focuses on managing distinct uncertain and unstructured large data sets using an efficient and secure clustering technique. This technique enhances data security using SDES encryption technique. The low space complexity and time complexity make it very suitable for large-scale data sets. Apart from that, it enhances the data privacy as well as integrity, and reduces data loss by protecting data from distinct security attacks and system errors [14]. Consequently, this technique can be applied for multidimensional large-scale data sets. The proposed technique also includes an efficient and low complex data searching scheme from the remote cloud or storage system which in turn reduces the data loss. The particular objectives of this research work are as follows:

- Developing an efficient clustering algorithm to offer dimensionality for the large-scale data sets; especially for the large-scale uncertain and unstructured data.
- Minimizing data loss by reducing distinct security attacks and transmission errors, and enhancing data privacy as well as the data integrity.
- Enhancing the data searching capacity to search from the remote cloud or storage system data with a low computational complexity.

The paper is organized as follows: Sect. 2 contains the background study to visit the existing security issues in managing uncertain and unstructured using distinct clustering algorithms. The background study section further examines the existing clustering techniques in terms of their strengths and weakness to explore the research gap. In Sect. 3, we propose an efficient clustering technique to address the

current limitations of a distinct clustering algorithm in managing uncertain and unstructured big data files. Section 4 concludes the effectiveness of the proposed work.

2 Literature Review

With the rapid uses of big data in our real-life applications, the security and management of different dimensional big data have become an important aspect of research. From the numerous research over the period, it has been observed that the clustering technique is one of the fruitful schemes to process as well as manage the big large data files of different dimensions. In this section, we review the strengths and weaknesses of existing clustering technique and diverse techniques for managing uncertain as well as unstructured big data. The literature study in this section further explores the strengths and weaknesses of various existing technique to manage multidimensional big data files.

The expanding utilization of online devices produces a huge amount of information in real time. The authors [10] have used this information to bolster reflecting abnormal state perspectives of data about the gathering created with foreseen designs symbolizing the conduct of strong gatherings.

Mehul [8] associated the efficiencies of HBase and MySQL for various arbitrary Read and Write processes. Rendering author, HBase produces higher efficiencies in comparison to MySQL during the handling of distributed file system. However, MySQL is showing better performances during complex query-based data handling, whereas HBase offers efficiencies for handling diverse formats of structured, unstructured and semi-structured large input datasets.

A comparison has been done by [14] among distinct NoSQL and SQL databases according to their ACID (Atomicity, Consistency, Isolation, Durability) and the BASE (Basic Availability, Soft state, Eventual consistency) characteristics. The work says that the SQL and the NoSQL databases are offering some common properties; however, their efficiencies are not similar in some of the specified circumstances. Hence, according to the authors, these circumstances are indicating that they cannot be identical for deciding any type of issue but one shall fairly choice between the two types of databases.

Andreu et al. [12] proposed an updated information segment procedure for Smart Grids by means of bunching information streams. This work acquainted an unconfirmed learning method with the enhancement in the execution of information stockpiling in Smart Grids. This procedure improved the developed classifier plan for bunching eXtended Classifier System for clustering (XCSc) calculation to display a crossbreed conspire that consolidates information duplication and isolating methodologies utilizing a web grouping technique. The proposed form of

XCSc makes it suitable for composite online surroundings as the proposed variant of XCSc can take in the right number of groups denied of accepting any from the earlier number of them in an online environment. The enhanced module makes XCSc equipped for acclimatizing itself to beforehand obscure examples regularly found at the Smart Grids stockpiling layer. The upgraded XCSc online structure is fit for overseeing extensive information sets helpfully with a genuinely ease regarding computational exertion. But it may require the client to legitimately set a few parameters to acquire precise results. Regularly, this is not an unimportant assignment and requires a specific level of ability in this sort of frameworks. Apart from that, this strategy cannot secure information uprightness.

Amrit et al. [9] proposed a Memory Utilization procedure for Hadoop group utilizing HDFS and MapReduce. In this exploration work, the creators investigated the variables like a measure of time spent by the maps and decreases, diverse memory utilizations by the mappers and the reducers for capacity and preparing of the information on a Hadoop bunch. But this strategy is not reasonable for complex time arrangement information and the multifaceted nature of this calculation is similarly high which is not ideal for the complex huge information applications.

Zhen et al. [5] utilized portraying and sub-setting strategy for Big Data Workloads. In this exploration work, the standard segment investigation (PCA) has been utilized to perceive the most key elements from 45 measurements to delineate huge information workloads from Big Data Bench, a comprehensive huge information benchmark suite. A grouping procedure has likewise been connected in this examination work utilizing the rule modules got from the PCA to research the resemblance among enormous information workloads, and the significance of including diverse programming stacks for huge information benchmarking has been confirmed. In this work, the measurable examinations on met measurements uncover that product stacks have vital impacts on workload conduct, and this impression is better than the calculations dynamic in client accommodation code. However, this strategy does not bolster a huge quantity of benchmark information. Besides that, this system is not reasonable for the multidimensional enormous information records.

Shen et al. [16] anticipated an execution checking method for vehicle suspension framework through fluffy positivistic C-implies grouping relying on the accelerometer estimations. Their proposed method can deal with the capacity weakening of springs just by tolerating the estimations of accelerometers fixed on the four corners of vehicle suspensions. It processes the quantity of groups first, recognizing blunders by FPCM and deficiency lines at second, and assuring the root elements finally. This suspension plan is exceptionally perplexing, and an extensive number of estimation is used to assemble framework demonstrate and not helpful for taking care of a vast number of vehicles.

Eric et al. [15] proposed a typical reason structure, Petuum that systematically talks information and model-parallel difficulties in huge scale machine learning, by seeing that endless machine learning projects are fundamentally advancement driven and admit blunder tolerant, iterative-focalized arrangements. The proposed Petuum offers machine learning tutors with a machine learning library and machine learning programming stage, achieved with overseeing big data and big machine learning models which show that is humble with specific solicitations, while running on fair-minded bunch sizes (10–100 machines). Nonetheless, Petuum is still similarly an undeveloped structure in contrast with Hadoop and Spark, and insufficiencies the consequent: flaw recovery from inadequate project state, capacity to change hold use on-the-fly in running occupations, booking employments for complex clients (multi-tenure), a unified information fringe that precisely blends with databases and spread file frameworks, and procurement for conveying scripting dialects, for example, Python.

An inclusive review has been presented by Diego et al. [13] which is concerning the efficiencies of different databases according to burden detecting data created by performance management tools. This study displays that the Cassandra imposts progressive latency time during both read as well as write processes. HBase needs fewer write latency but offers great read latency time. Consequently, HBase shows better capability for handling distributed large input files rather than the Cassandra database.

Biswas [3] proposed a special type of distributed system called "Atrain Distributed System" (ADS) which is very suitable for processing big data using the heterogeneous data structures r-atrain or the homogeneous data structure r-train. A simple "Atrain Distributed System" is called an uni-tier ADS. The "Multi-tier Atrain Distributed System" is an extension of the uni-tier ADS. The ADS is scalable up to any extent as many times as required. Two new types of network topologies are defined for ADS called by "multi-horse cart" topology and "cycle" topology which can support increasing volume of big data. Where r-atrain and r-train data structures are introduced for the processing of big data, the data structures "heterogeneous data structure MA" and "homogeneous data structure MT" are introduced for the processing of big data including temporal big data too.

In the next section, we summarize the problem precisely and present its solution.

3 Secure Clustering Technique

From the existing literature, it can be observed that the existing clustering technique or other current techniques are incapable of managing the uncertain and unstructured as well as multidimensional big data. These inadequacies of existing techniques for managing such big data considered in our work here by proposing a secure clustering technique which additionally offers the protection against data and higher data privacy.

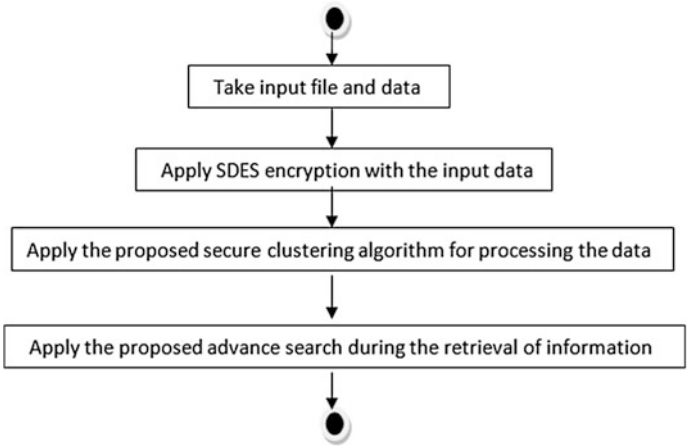


Fig. 1 The proposed model of secure clustering technique

The proposed model has basically three steps. In the first step, the SDES encryption algorithm has been applied to make the information secure, details of SDES can be studied in the work of Pronpitag et al. [11]. After that, the clustering algorithm is applied to process and store the encrypted data. At the time of retrieval of encrypted data, a low complex search algorithm has been applied. The steps are shown in Fig. 1.

From Fig. 1, the corresponding algorithm Secure Clustering Technique Algorithm (**SCTA**) is designed which is presented below.

It can be seen that, in our technique, we have the following for our **SCTA** algorithm:

Our inputs are: the number of Cluster is N, the Data Set S, the Data Objects $D = \{d_1, d_2, d_3, \ldots, d_m\}$, and also the set of attributes at any point are given by $A = \{a_1, a_2, a_3, \ldots a_m\}$ where $(0 \leq n, m < \infty$ and $n \leq m)$.

Clearly, our Output will be a set of Cluster N.

```
Algorithm :   SCTA
1. Draw the multiple subsamples { s1, s2, s3,… sm}
   from the original data sets S.
2. Take the middle point of each set as centroid
3. Compute the distance between each data point and all the initial centroid.
4. For each point, find the closest centroid and assigned to the nearest
   cluster.
5. Choose the minimum distance from the cluster to the centroid.
6. Apply steps (1-5) on data sets for cluster N
7. Combine two cluster into single cluster.
8.  Calculate the new cluster center for the combined cluster until the
   number of clusters becomes N.
```

Basically, in our search technique, two types of search techniques have been integrated. Initially, in the normal search technique searches the encrypted text with

the help of concatenated search key. If the search gets failed then the hybrid search technique needs to apply for searching. In this advance search method, the key value for the search technique is converted into cryptic text with the use of SDES encryption technique and then searching is performed. Thus, the proposed secure clustering technique can execute uncertain and unstructured as well as the multi-dimensional big data files.

4 Conclusion

In this research work, we have investigated the existing techniques for storage and managing multidimensional big data, especially the uncertain and unstructured big data files in terms of their strengths and shortcomings. We have developed a secure clustering technique which can manage multidimensional big data files as well as uncertain and unstructured big data, which fulfills our first objective. The corresponding algorithm SCTA is also designed.

Then, the proposed technique includes SDES encryption technique which protects the data from various system errors as well as security attacks with the minimum execution complexities. Consequently, this encryption technique makes the proposed technique enabled to protect the data loss. Apart from that, the proposed technique includes an advanced clustering technique for managing and processing a large amount of multidimensional big data, especially uncertain and unstructured data. This clustering technique also helps to reduce the data loss due to data overhead. Thus the proposed technique satisfies our second objective. An advanced searching technique has been incorporated with the proposed technique to make the data retrieval process faster and more efficient with the minimal computational complexity, which accomplishes our third objective of this research work.

In our future work, implementation of the proposed secure clustering technique will be implemented and the result will be analyzed in different aspects. Depending on the performances in various security aspects, further modifications in the proposed secure clustering technique will be made. Apart from that, the capacity of our proposed technique in a different environment will be tested with standard multi-dimensional, especially various unstructured and uncertain big data sets.

References

1. Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U.: Challenges and opportunities with Big Data. Cyber Center Technical Report 2011-1, Purdue University, January 1, 2011
2. Alomari, M.A., Khairulmizam, S.: A framework for GPU-accelerated AES-XTS encryption in mobile devices. In: TENCON 2011 IEEE Region 10 Conference (IEEE), pp. 144–48 (2011)
3. Biswas, R.: Atrain distributed system (ADS): an infinitely scalable architecture for processing big data of any 4Vs. In: Acharjya, D.P., Dehuri, S., Sanyal, S. (eds.) Computational

Intelligence for Big Data Analysis Frontier Advances and Applications, pp. 11–53. Springer, Switzerland (2015)
4. Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A.A., Lally, A., Murdock, J.W., Nyberg, E., Prager, J., Schlaefer, N., Welty, C.: Building Watson: an overview of the DeepQA project. In: AI Magazine, Fall, pp. 59–79 (2010)
5. Jia, Z., Zhan, J., Wang, L., Han, R., McKee, S.A., Yang, Q., Luo, C., Li, J.: Characterizing and subsetting Big Data workloads. In: IEEE International Symposium on Workload Characterization (IISWC) Raleigh, pp. 191–201 (2014)
6. Kang, X., Xu, X., Peng, A., Zeng, W.: Scalable lossy compression for pixel-value encrypted images. In: Data Compression Conference (DCC), p. 400. IEEE (2012)
7. Liu, C.-H., Ji, J.-S., Liu, Z.-L.: Implementation of DES encryption arithmetic based on FPGA. In: AASRI Procedia, vol. 5, pp. 209–213. Elsevier (2013)
8. Mehul, N.V.: Hadoop-HBase for large-scale data. In: International Conference on Computer Science and Network Technology (ICCSNT), pp. 601–605, Harbin (2011)
9. Pal, A., Agrawal, S.: An experimental approach towards Big Data for analyzing memory utilization on a Hadoop cluster using HDFS and MapReduce. In: First International Conference on Networks and Soft Computing (ICNSC), Guntur (2014)
10. Perera, D., Kay, J., Koprinska, I., Yacef, K., Zaiane, O.R.: Clustering and sequential pattern mining of online collaborative learning data. IEEE Trans. Knowl. Data Eng. **21**(6), 759–772 (2009)
11. Puangpronpitag, S., Kasabai, P., Pansa, D.: An enhancement of the SDP security description (SDES) for key protection. In: 9th International Conference on Electrical Engineering/ Electronics, Computer, Telecommunication and Information Technology, ECTI-CON, pp. 1–4. IEEE (2012)
12. Sancho-Asensio, A., Navarro, J., Arrieta-Salinas, I., Armendáriz-Íñigo, J.E., Jiménez-Ruano, V., Zaballos, A., Golobardes, E.: Improving data partition schemes in Smart Grids via clustering data streams. Expert Syst. Appl. **41**, 5832–5842
13. Serrano, D., Han, D., Stroulia, E.: From relations to multi-dimensional maps: towards an SQL-to-HBase transformation methodology. In: IEEE 8th International Conference on Cloud Computing, New York City, NY, pp. 81–89 (2015)
14. Tudorica, B.G., Bucur, C.: A comparison between several NoSQL databases with comments and notes. In: RoEduNet International Conference 10th Edition: Networking in Education and Research, Iasi, pp. 1–5 (2011)
15. Xing, E.P., Ho, Q., Dai, W., Kim, J.K., Wei, J., Lee, S., Zheng, X., Xie, P., Kumar, A., Yu, Y.: Petuum: a new platform for distributed machine learning on Big Data. IEEE Trans. Big Data **1**(2), 49–67 (2015)
16. Yin, S., Huang, Z.: Performance monitoring for vehicle suspension system via fuzzy positivistic C-means clustering based on accelerometer measurements. IEEE/ASME Trans. Mechatron. **20**(5), 2613–2620 (2015)

Reducing Search Space in Big Data Mining

Surabhi Kumari, V. G. Sathve and Savita K. Shetty

Abstract Data in many real-life applications are voluminous and riddled with uncertainty. Big data represents a very large amount of data having more varied and complex structure with the challenges of searching, storing, analyzing, and visualizing for further processes or results. Thus, it becomes essential to develop and apply sophisticated algorithms to process this enormous amount of complex data. With the advent of technology, data is generated at a staggering rate, and spans unstructured and semi-structured data along with static data. With the generation of such huge data comes various challenges, One of them is Searching and the reduction of Search Space. Presently, Search Space for users is very large and sometimes may contain data irrelevant to their search. In this paper, we discuss reduction of search space for text-based information using a single machine clustering algorithm. Clustering is done before applying any searching algorithm as it partitions the data into various clusters and limits the search space to a particular cluster rather than searching in the whole dataset. This enables easy processing of huge quantities of data and for that we use Fuzzy-c means clustering algorithm. Fuzzy-c clustering method provides a better and more efficient method to cluster these patterns by using degree of membership. Beyond this, an efficient search algorithm is implemented called, the Lucene search algorithm. Lucene search performs indexing on input data and then use that indexed data to perform searching. Lucene search algorithm first analyzes the input data and stores the input in a data structure called inverted index which facilitates efficient retrieval of information. Although Lucene only works with text files, it can also be modified to be used for other kinds of data that can be converted into text files. Thus, in this paper, we bring out the implementation details of aforesaid algorithms and the

S. Kumari (✉) · V. G. Sathve · S. K. Shetty
Department of Information Science & Engineering,
M S Ramaiah Institute of Technology, Bangalore, Karnataka, India
e-mail: surbhik209@gmail.com

V. G. Sathve
e-mail: sath.hp@gmail.com

S. K. Shetty
e-mail: savita_ks1@msrit.edu

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_46

results recorded show a significant improvement in performance parameters of the search.

Keywords Fuzzy-c means · Lucene · Big data · Big data analytics
Indexing · Lucene search

1 Introduction

Data mining [1] is a technique which can be used to search for information which is not known previously and can be useful from the input dataset. Data mining algorithms can be applied to different real-life applications such as grouping similar people based on similar properties (Clustering) and categorization based on old records (Classification). With the progress of technology, large quantity of valuable data in the varied streams of banking, marketing, data from social networking sites, telecommunication, medical are generated in many real-life applications. This brings us to the new age of *Big data* [2]. Big data refers to interesting high velocity, high valued data with volume beyond the capabilities of traditional data mining techniques to capture, store and process. Processing this large amount of data cannot be done within a tolerable elapsed time. Therefore, introduction of modern processing techniques is required so that complexity of those large volume data can be reduced and can be handled easily so that some meaningful information can be retrieved. This encourages the research and developments in *Big data analytics.*

To handle *Big data* and to overcome the challenges faced in processing that large data, use of Clustering is introduced which is basically used to partition the data into various groups based on some similar features which help us to focus on a particular subset of data at a time which may later be useful for various other operations like searching, storing, etc.

Over past few years, many algorithms are introduced which are useful in partitioning of data (clustering algorithms) and *Fuzzy-c means* [3, 4] is one of those algorithms. Presence of various clustering algorithms like fuzzy-c means, k-means, hierarchical clustering, etc., leads to the idea of clustering analysis which is a way of putting together a set of data points such that those in the same group (i.e., cluster) are more similar or related to each other than the ones in the other group (i.e., cluster). It is the main task of exploratory data mining; this technique is commonly used for statistical data analysis and in many other fields like machine learning, pattern recognition, information retrieval and data compression.

Leung and Kyle [3] and Ertöz et al. [4] *Fuzzy-c* means algorithm is introduced by Bezdek and is an extension of Hard C-Mean clustering method. This algorithm overcomes the shortcomings of other previous clustering algorithms like K-means which provides a hard boundary in clustering.

This algorithm performs analysis on the basis of distance between various input data points. First random cluster centers are chosen. The distance between data points and cluster centers are computed. Data points belong to that cluster whose

cluster center is the nearest. The degree of membership for every data item in each cluster is the deciding criteria to insert that data item into its appropriate cluster is the coefficient that tells the membership degree (u_{ij}) of every data point belonging to the kth cluster.

$$w_{ij}^{m} = \frac{1}{\sum_{k=1}^{c} \left(\frac{||X_i - c_j||}{||X_i - c_k||}\right)^{\frac{2}{m-1}}} \tag{1}$$

where W_{ij}^{m}:- membership values & m:- fuzzifier.

If a data point belongs to more than one cluster, then the fuzzifier(m) is used to determine its existence. It is an user-defined unit lying between 0 and 1. Fuzzifier determines the level of cluster fuzziness. A large m results in smaller member-ships and hence, fuzzier clusters. It is commonly set to 2.

The main objective of the fuzzy clustering algorithm is to create clusters such that there is a maximum similarity with the data points in the same cluster and there is minimum similarity with the data points in different clusters.

Since big data represents an interesting large-valued, unstructured, and high-volume data, there are various challenges that are faced in the processing of that high volume data. Challenges in big data mining includes analysis, capture, search, storage, transfer, visualization.

Puthal et al. [5] and Zhang and Timothy [6] *Lucene* is very popular and fast search library used in Java-based application to facilitate document searching in any kind of application in a very simple and efficient way. It can be used in any application to add search capability to it. It is scalable and high-performance library used to index and search virtually any kind of text. Lucene algorithm works as a very important part of any search application and provides the vital operations pertaining to indexing and searching is to be performed on the clustered set of data. The clustered data is taken and documents need to be built from the clustered data content which can be understood and interpreted by the algorithm easily. Before indexing process to start, documents need to be analyzed for the part of the text which needs to be indexed or the text which is the candidate to be indexed. After the documents are built and analyzed, they should be indexed so that documents can be retrieved based on certain keys instead of the whole content of the document. This way an inverted index table is generated for the documents which consist of a mapping between different words and the documents consisting of those words along with their frequencies which makes the document retrieval a lot easier. Once the database of indexes is created, the application is ready to make any kind of search and to facilitate user to make a search, the application must provide the user a means to enter text and start the search process. The user input is then used to prepare a query object using the text input which can then be used to get the relevant details. Using query object, index database is then checked to get the relevant details and the content documents.

2 Existing Work

2.1 Extraction of Frequent Patterns

In Lahcen and Mouline [1] the existing model permit the user to enter SQL-based queries to specify their constraints for extraction of frequent patterns. Only the pattern matching the user-based queries are extracted. By this means, unwanted calculations for mining those undesired frequent patterns can be avoided.

Constraint C1 $\equiv$ min (Y.age) $>=$ 18 so that the minimum age for voting in all census records for a unit Y is at least 18. Constraint C2 $\equiv$ Y. Height $\geq$ 170 cm so that the height of every attribute in Y is at least 170 cm. Constraint C3 $\equiv$ Y. Place = London which shows the interest of the user in every event taking place in a particular location.

Constraits specified by the user can be generally classified into classes which overlap depending on the properties they own.

2.2 The MapReduce Programming Model

MapReduce [1] defines a high-level programming model which can be used to process a large amount of data or any kind of huge dataset. MapReduce basically uses the concept of parallel and distributed computing systems where multiple systems are connected together by means of some kind of network so as to achieve the same goal so that computational capabilities of any system can be improved by processing the data at distributed systems. So, parallel data processing can be done at multiple locations using this model which reduces time and effort also.

Two main methods involved in case of this model are Map and Reduce. Map() function finds a match between keys and values pairs and returns as output a set of (key, value) pairs. Its main function is filtering and sorting. Unwanted data is removed and the rest is sorted in a particular order. Then keys are matched with its appropriate values and all sets of such pairs are returned. Example: name - > age where name is the key and age is the value. Reduce function combines all similar key value pairs to get a single key-value pair. Its main function is to return a summary of the whole processing part.

Example:- for counting number of times a word occurs in a text file.

a- > 1,the- > 1,a-1,a- > 1,hello- > 2 are the key-value pairs generated during map

a- > 3,hello- > 2,the- > 1 are the key-value pairs generated during reduce.

The value of the same key is combined in the reduce function. Thus, reducing the number of key-value pairs.

Various examples of this application may involve the construction of reverse web link graph, count of URL access frequency and the word count for a document.

3 Proposed Algorithms for Clustering and Performing Search on that Clustered Data

Proposed method for reducing the search space in the huge dataset tries to first partition the raw input textual data based on the frequency of certain related words and then applying searching algorithm on that partitioned data instead of wasting time performing searching operations on the whole large dataset which takes a lot of time processing and search the data. We use Fuzzy-c means algorithm for clustering the dataset and Lucene search algorithm for performing efficient search and the algorithms are explained as follows:

3.1 Fuzzy-C Means Clustering Algorithm

Fuzzy-c means [4] algorithm provides a very efficient clustering method as Fuzzy clustering assigns membership levels, and then use those membership levels to assign data elements to various clusters. In fuzzy clustering, every data point is associated with membership levels with each cluster and every data point can belong to two or more clusters. These membership levels estimate the strength of the association between that data element and a particular cluster.

Given a finite set of data, the algorithm returns a list of cluster centers and a partition matrix, where each element in the partition matrix tells the degree to which data point belongs to cluster.

Properties of fuzzy-c means are the following:

1. Only the degree of membership of a particular data point associated with a particular cluster is calculated the absolute membership of a data point associated with a cluster is not evaluated.
2. Fuzzy-c clusters represent a soft boundary between multiple clusters rather than having hard boundary between two clusters so a data point can belong to more than one cluster having some membership value in every cluster.

FCM aims to minimize an objective function:

$$\underset{C}{argmin} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}^{m} \left|\left| x_i - c_j \right|\right|^2$$

where w_{ij} represents the membership value of a data point in a cluster & is given by

$$w_{ij}^{m} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\left|\left| X_i - C_j \right|\right|}{\left|\left| X_i - C_k \right|\right|} \right)^{\frac{2}{m-1}}} \tag{3}$$

m:- fuzzifier, fuzzifier determines the level of cluster fuzziness. It is commonly set to 2.

The best part about using fuzzy-c means algorithm is that gives best result for overlapped data set and data point is assigned membership with respect to each cluster result of which data point may belong to more than one cluster at the cost of long computation time and it gives better results for lower m value at the expense of more no of iterations.

3.2 Lucene Search Algorithm

Lucene [5] is a fast search library used in Java-based application to facilitate document searching in any kind of application in a very simple and efficient way. In a nutshell, when Lucene indexes a document it breaks it down into a number of terms. It then stores the terms into index files where each term is associated with the document that contains it. We can consider it like a hash table. Terms are generated using an analyzer which stems each word to a root form. When a query is issued it is processed through the same analyzer that was used before to build the index and then used to search for the matching terms in the index. This provides a list of documents that match the query.

To analyze Lucene's index file structure, we need to understand the concept of inverted index structure. In this an inside out arrangement of documents is used where terms take center stage. Each term is then associated with the documents that contain it in inverted index concept, whereas in a forwarding index, documents are analyzed to generate a list of terms it contains. Lucene, hence uses an inverted index structure.

Lucene search algorithm has been widely implemented by well-known organizations. For e.g., it provides searching capabilities for the Eclipse help system, MIT's OpenCourseWare, etc.

4 System Implementation

This paper employs the use of fuzzy-c means algorithm for partitioning of data into various clusters and after that Lucene search algorithm is applied so as to increase the efficiency of searching. Lucene's primary goal is to facilitate information retrieval. Parsing and analysis of data are done instead of indexing and searching data so that best search results are obtained. Beyond this, if the input data is clustered and passed to the Lucene search in the initial state then it gives a better result in form of the searching efficiency.

4.1 Clustering of Data

For clustering, Fuzzy-c means provides a model where data point is assigned to a particular cluster based on the degree of membership attached to a particular document or the data point. For the calculation of the membership degree, distance between data points and the cluster center is taken into consideration. A document is assigned a cluster to which it has the highest level of membership. But, since distance measure is the main feature of fuzzy clustering, it does not work well for data in textual format. Therefore, in our case, we design a modified version of fuzzy-c algorithm which works for textual data as fuzzy-c is mostly limited to numeric-based data. In this algorithm, we use the most probable frequent words from each cluster for assigning the appropriate clusters in place of distance parameters used in fuzzy-c means algorithm. In case one data belongs to multiple clusters, this conflict is resolved by checking for the frequency of those words in each cluster for the respective documents and the cluster having the maximum frequency is assigned. So, in this case membership level is defined by the frequency or presence of the predefined set of probable words in various clusters.

Implementation of this modified fuzzy-c can be carried out in the following way:

(1) Raw input is collected in the textual form of data.
(2) No of clusters initialized.
(3) File path for all clusters is defined.
(4) Input file directory path is defined.
(5) Most probable frequent words for each cluster is specified and stored so that it can be accessed and used while assigning of clusters.
(6) List all the files in the main input file directory.
(7) For every file stored in input directory, match the words in the file with the probable frequent words of cluster. If those words are present in the file, assign those files to that particular cluster.
(8) Above steps are repeated till all files are assigned to some cluster.
(9) There may be a case where one file may be assigned to more than one cluster. That creates conflict in cluster contents.
(10) In that case, maximum frequency of those words is calculated and then that file is assigned to a cluster having the maximum frequency of that particular word.
(11) After the assignment of all the files is complete, file path for each cluster is stored so that it can be accessed while searching is to be done.

4.2 Searching Algorithm

Zhang and Havens [6] for searching, we use the concept of Lucene search algorithm where the input data is first built into documents, analyzed before indexing is to be

applied on that. Lucene primary goal is to facilitate information retrieval. Indexing and searching steps have to be combined with parsing and analysis of input to achieve the best search results. Information retrieval has several sequences of steps.

Lucene Development Kit version is LUCENE-CORE-3.6.2. JAR, it also requires Java runtime environment above JDK1.6 version and JAR package must be imported into Eclipse.

Various modules involved in the searching procedure are:-

(1) **Preprocessing Module**:- Before applying Lucene, data need to be preprocessed so that information retrieval becomes easier and faster. We use clustering as a preprocessing step. We divide the whole dataset into various partitions so that file size can be decreased. For this, various documents in the input data are divided into clusters based on the presence of some probable words. These clusters can then be used while searching as based on the user input to be searched clusters are located, cluster path is accessed and then documents inside those clusters are only searched. This way, searching is restricted to a subset of the input data and time is saved.
(2) **Indexing Module**:- Indexing process is one very important functionality provided by Lucene. IndexWriter is the core component of the indexing process. Document added to IndexWriter gets analyzed using analyzer and then indexes are created as required and then those indexes are stored in the directory where it can be stored as well as updated. Indexwriter is used just to create and update indexes, it cannot be used to read indexes.

Indexing process involves various steps like [7]:-

Creation of documents:- first a method is created to get documents from the text files. Various types of Fields are created which represents key value pairs with keys as names and values as contents to be indexed. Analysis of documents is to be done and for that a set of stop words are defined like a, am, the, is etc. which doesn't need analysis since it is not required in the searching process. After the analysis of data is done, newly created fields are added to document objects.

Creation of IndexWriter:- IndexWriter acts as a core component which creates/updates indexes during indexing process. In this process, first an object of indexwriter class is created. A directory is created which points to the location where indexes are to be stored. In the indexWriter object analyzer information is provided so that documents can be analyzed and indexes are to be created.

Now for indexing to start, first the file path is retrieved and then documents are accessed using *getDocument(file)* method and then the document is added to indexWriter.

(3) **Searching Module:-** This module implements the main objective of Lucene search where IndexSearcher is used which is the one of the important components of the searching process. First Directory(s) containing indexes is

created and then passed to IndexSearcher which can open the Directory using IndexReader. User input is taken for the term to be searched and for that a Query is created and search is made using IndexSearcher by passing the Query to the searcher. IndexSearcher returns a TopDocs object which contains the search details along with document ID(s) of the Document which is the result of the search operation.

QueryParser:-

- QueryParser object created parses the query of the user to the format understandable.
- QueryParser object is initialized with analyzer having an appropriate version number and index name of the query.

IndexSearcher:-

- IndexSearcher object is created and initialized with Lucene directory.
- It is used to access the index table stored in the Lucene directory so that searching will be done only on the index table to make searching efficient and faster.

Search operation:-

- It will take the input from the user in the form of queries and is passed to the queryParser which will parse the query to a format understandable by the queryParser.
- Search method is called to give the results of the search operation.

Retrieving the document
Steps for Lucene search implementation:-

Step 1: Initialize the indexer to point to the index directory.
Step 2: All files stored in folder are sent for indexing if they have no IOExceptions.
Step 3: Documents are created by initializing fields of the files like file name, path to the document.
Step 4: File contents are always analyzed before creating the document to remove stop words.
Step 5: Document is written into IndexWriter and number of documents are returned.
Step 6: A query parser is created which parses the query provided by a user to a format understood by Lucene.
Step 7: An indexSearcher points to the index directory to do searching.
Step 8: Top documents that match the query are returned.
Step 9: Searcher and indexer are closed.

Table 1 Comparing search operation with respect to clustering

File size	Searching time (ms)		Indexing time (ms)	
	With clustering	Without clustering	With clustering	Without clustering
1.25 GB	28	35	469	227896
2.1 GB	75	90	116830	412140

4.3 Performance Comparison of Searching with Respect to Clustering

So, indexing time depends on the size of data. So, we see that as the file size increases, time taken for indexing as well as searching increases. So, clustering plays a very important role here to keep the file size smaller so that searching and indexing time can be reduced. Searching time mostly depends on the indexed data for how well data is indexed because searching is performed on the indexed data. So, in some cases, searching time may be less for un-clustered data (Table 1).

5 Conclusion

Searching data is one of the biggest problems faced in Big Data analysis. This paper resolves to solve this problem by clustering the data, then applying Lucene search algorithm for greater speed and efficiency. The concept of indexing documents before searching speeds up the searching process and makes information retrieval faster but when it comes to very large volume of data indexing process takes time and in that case clustering plays a very important role as it divides the data into various clusters of smaller sizes thus making the indexing and searching much more efficient. Thus, search space of big data can be reduced by clustering the data and then applying search algorithm on that particular cluster.

References

1. Lahcen, A.A., Mouline, S.: (LRIT, Unit associated to CNRST URAC 29, Mohammed V University, Rabat, Morocco)
2. Zerhari, B.: Big Data Clustering: Algorithms and Challenges
3. Leung, C.K.-S., MacKinnon, R.K.: Reducing search space for Big data mining for interesting pattern from uncertain data. Fan Jiang published at IEEE International Congress on Big data (2014)
4. Ertöz, L., Steinbach, M., Kum, V.: An Approach Towards The Shared Nearest Neighbor (SNN) Clustering Algorithm. University of Minnesota
5. Puthal, D., Nepal, S., Paris, C., Ranjan, R., Chen, J.: Efficient Algorithms for Social Network Coverage and Reach. Published by IEEE

6. Zhang, Z., Havens, T.C.: Scalable Approximation of Kernel Fuzzy c-Means. Published by IEEE
7. https://www.javaranch.com/journal/2004/04/Lucene.html
8. Suganya, R., Shanthi, R.: Fuzzy-c means algorithm–a review. De-partment of CS Dr. SNS Rajalakshmi College of Arts and Science. Int. J. Sci. Res. Publ. **2**, Nov 2012
9. Borhade, S.B.: Best search and retrieval performation evaluation with lucene indexing. Prof. Pankaj Agarkar Published at Multidisciplinary Journal of Research in Engineering & Technology
10. Ramaprasath, A., Srinivasan, A., Lung, C.-H.: Performance Optimization of Big Data in Mobile Networks. Published by IEEE
11. Gao, R., Li, D., Li, W., Dong, Y.: Application of Full Text Search Engine Based on Lucene
12. Performance Comparison of Fuzzy C Means with Respect to Other Clustering Algorithm by
13. Marmanis, H., Babenko, D., Greenwich, M.: Algorithms of the Intelligent web
14. https://www.lucenetutorial.com/techniques/indexing-databases.html
15. https://www.lucene.apache.org/

Justified Group Recommender Systems

Venkateswara Rao Kagita, Arun K. Pujari and Vineet Padmanabhan

Abstract *Justification* improves the reliability of a recommender system because it helps user/s understand the *reasoning* behind the recommendation. `Nearest neighbor style` and `influence style` are the common justification styles in a recommender system. Since both styles are constructed exclusively in light of user preferences on the item rather than *content* of an item, the recommendation cannot be adequately justified. Moreover, these justification styles are applicable for personal recommender systems rather than group recommender systems. In this paper, we introduce a novel justification style for group recommender systems having the structure "item `x` is recommended because *those who watch* `y, z,...` *that contain features* $\{g_i, g_j, \ldots\}$ *also watch* `x` *that contains* $\{g_k, g_l, \ldots\}$". Our justification style is based on *precedence mining* model, wherein the precedence probability of using an item by an active user is determined based on pairwise precedence relations between the items. We broaden this idea of precedence probability to accommodate the social influence factor. No past investigation deals with justified group recommender systems.

Keywords Group recommender systems · Justification style · Precedence mining

1 Introduction

Group Recommender Systems (GRS), are an extension of Personalized Recommender Systems (PRS), aim at recommending interesting items to a group of peo-

V. R. Kagita (✉) · A. K. Pujari · V. Padmanabhan
University of Hyderabad, Hyderabad, India
e-mail: 585venkat@gmail.com

A. K. Pujari
e-mail: akpcs@uohyd.ernet.in

V. Padmanabhan
e-mail: vineetcs@uohyd.ernet.in

A. K. Pujari
Central University of Rajasthan, Ajmer, India

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_47

ple. It has been increasingly popular nowadays due to its application in many domains. Researchers have explored different ways to extend PRS to GRS, (1) *Merging Scores*—Aggregates all the group members score, for each candidate item for a recommendation, by obtaining score using each profile separately [2, 6, 10, 15]. (2) *Merging Profiles*—Combines individual profiles by combining individual ratings from the history of consumption [13, 15]. (3) *Merging Recommendation*—The group recommendation is obtained by combining individuals recommendation [1, 4]. All these strategies have limitations as pointed out by [10] in a sense that unforeseen relationship between the individuals and their gathering have not been focused. For instance, it is possible that individuals can alter their personal preferences to suit those of other gathering people who they know for more enjoyment from the group or to consume diverse sorts of things. Moreover, if we consider group-level preferences to be the intersection purpose of all the group members solitary choices, then the common interests among every one of the individuals could be somewhat limited which in turn could result in data sparseness.

To overcome the above limitations, recently we propose a *virtual user* [7–9] for group recommendation based on precedence relations and demonstrated that it is better than various group modeling strategies. Nevertheless, none of the works mentioned till now have talked about giving an explanation/justification for a group recommender system. To quote briefly, justification in recommender system is an explanation given to a user to justify the recommendation. Justified recommendation provides credibility to a recommender system and gains customer trust and acceptance [14].

In this paper, we propose a Justified Group Recommender Systems (JGRS) and we consider most recent and successful group recommender system [9] as an underlying algorithm. The main idea is to provide justification/explanation for the recommendations made by the system. No early attempt has done to develop a justified *group* recommender system though there are some works on justified *personal* recommender systems. One of the downsides of precedence mining model is that they ignore the *new* items as they give *scores* based on other users consumption. Therefore, we need to apply precedence mining on the set of features of an item which in turn makes an item to be represented as a *vector* instead of *item-id's* as done in traditional precedence mining approach. To consider the influence of certain specific attributes in the overall recommendation process we also examine a concept called *social influence factor*. For the justification part, we construct a genre preference matrix from vector precedence information and calculate *score* using *top-I* precedence probabilities for each item.

The specific contributions of our work are as follows:

1. We extended the precedence mining based group recommender model developed in [9] to accommodate the concept of *justification* in recommender systems.
2. *Social influence* factor is taken into account while giving recommendation and the experimental results show that quality of recommendation is improved with the addition of social influence.

3. We proposed a new measure called *Justification rate* to measure the justification quality.

Rest of the paper is organized as follows. In Sect. 2, techniques related to extending precedence mining for group recommender systems are discussed. Social influence model is given in Sect. 3. Section 4 describes the proposed *justified recommendation model* and Sect. 5 is comprised of experimental analysis.

2 Precedence Mining-Based GRS

The idea behind precedence mining model is to learn the temporal patterns in the usage for the recommendation. For instance, a user who viewed an item x may choose to view an item y but not vice versa. Precedence mining-based recommender system [12] extracts precedence statistics from the customer usage records. It uses those statistics and active (target) user history to compute the score for a candidate item o_c by an active user u_a as given below.

$$Score(o_c, u_a) = \frac{support(o_c)}{n} \times \prod_{o_l \in O_a}^{(I)} \frac{precedence(o_l, o_c)}{support(o_c)} \tag{1}$$

where $support(o_i)$ is the number of users consumed item o_c, $precedence$ (o_i, o_j) is the number of customers having consumed item o_i before item o_j and O_a is the set of items consumed by an active user u_a. $\prod^{(I)}$ means product of highest I values. After computing the score for each candidate item, PRS recommends top k items to an active user u_a. On the other hand, GRS has to calculate the group G score for every candidate item. The simplest way to extend above model to GRS is merging scores strategy. That is, calculate the score by every member of the group and then aggregate. As mention in the introduction section, it does not preserve the group relations well. To overcome this problem, in our previous work [9], we propose a virtual user strategy. We briefly review it in the next subsection.

2.1 Virtual User Strategy

Virtual user strategy represents the whole group with a single virtual user. We propose two different approaches to build a virtual user profile, (1) threshold-based virtual user, (2) weighted virtual user. In this work, we use weighted virtual user. In the case of a weighted virtual user, we include every group item[1] in the

[1] An item is said to be a group item if at least one user in the group consumes that item.

virtual user profile with associated weight. Weight of a group item o_i is defined as $weight(o_i, G) = \sum_{u_a \in G} weight(o_i, u_a)/|G|$ and

$$weight(o_i, u_a) = \begin{cases} 1 & \text{if } o_i \in O_a \\ score(o_i, u_a) & otherwise. \end{cases}$$

After getting the virtual user profile, compute the score of each candidate item as follows:

$$score(o_c, G) = \frac{support(o_c)}{n} \times \prod_{o_l \in O_{v(G)}}^{(I)} weight(o_l, G) \times \frac{precedence(o_l, o_c)}{support(o_c)} \quad (2)$$

where $v(G)$ denotes virtual user of a group G.

3 Social Influence Model

Users decision could be influenced by many unseen factors behind items not just items identity. The factor could be the content of an item or a person associated with it. Hence, it is more appropriate check the temporal patterns of those factors rather on item-ids. For example, in movie recommendation, a user may like to watch comedy movie after horror film. Similarly, those who watch actor A movies may like to watch actor B movies [7]. In the present work, we use these two factors (1) feature vector of an item: we represent each item in the user profile with its decimal equivalent of a feature vector, and (2) actor of an item: We represent user profile as a sequence of actors instead of item ids. Generalizing this model to multiple other factors is trivial. Our thorough experimental analysis on benchmark dataset observed that quality of the recommendation had been improved (results related to this are reported somewhere else). User u_a's score for a movie/item o_c in this case can be calculated as in Eq. 3.

$$score(o_c, u_a) = \alpha.score_{vec}(o_c, u_a) + \beta.score_{actor}(o_c, u_a) \quad (3)$$

where $score_{vec}(o_i, u_j)$ is the score obtained by viewing user profile as a sequence of vectors and $score_{actor}(o_i, u_j)$ is the score obtained by viewing user profile as a sequence of actors rather than a sequence of items. α and β are constants that represent the importance of vectors(content) and actors(social influence). In our implementation, we have given equal importance to both content and social influence, In other words, $\alpha = \beta = 0.5$.

4 Justified Group Recommender System

4.1 Justification

Justification is a reason, fact or explanation that justifies the recommendation. Justification requires information that should be kept in mind when making a decision, that will be helpful while explaining why we have taken a decision of recommending a particular item. Providing justification along with the recommendation is so important because of the following reasons.

- The acknowledgment of a recommender framework is expanded when clients can understand the qualities and confinements of the suggestions. This can be accomplished when clients get, alongside a recommendation, the thinking behind it.
- It helps the user to understand the malicious/shilling attacks [11]. If an unjustified recommender system is under shilling attack, it is difficult for users to comprehend why they get undesirable suggestions.

4.2 Justification Style

Many real-time e-commerce sites also adopted justified recommedation model. For instance, Amazon[2] and Flipkart[3] adopted `nearest neighbor style` of justification [3], "customers who viewed product x also viewed product y". Another style is `influence style:` "Item x is recommended because you like item y" [14]. In both the styles, it becomes difficult for a user to understand the link between x and y. If we explain how these two items are related in detail, by capturing content relation and social influence relations then it would be a more acceptable style of justification. To this end, we introduce a keyword-based precedence relational style of justification, as follows. "Movie x is recommended for you because: (1) those who watches y, z,... that contain $\{g_1, g_2, \ldots\}$ also watches x that contains $\{g_i, g_j, \ldots\}$, (2) those who watches movies of a, b, ... also watches movies of hero c". 1 and 2 in above justification can be interchanged based on the importance of content/social influence.

4.3 Justified Recommendation Model

We use *social influence model* as described in Sect. 3 for building justification framework along with the recommendation. Equation 3 specifies how to calculate a score in social influence model. Where in $Score_{actor}(o_i, u_j)$ is the score obtained

[2] https://www.amazon.com/.

[3] https://www.flipkart.com/.

Table 1 Converting Vector precedence information to genre precedence information

	v_3	v_4
v_1	3	8
v_2	9	4

(a) VP

	g_1	g_2	g_3	g_4	g_5
g_1	11	8	3	8	0
g_2	0	0	0	0	0
g_3	11	8	3	8	0
g_4	0	0	0	0	0
g_5	13	4	9	4	0

(b) GP

by extending $profile(u_j)$ as a sequence of actors in case of movie recommendation. $Score_{vec}(o_i, u_j)$ is the score obtained by extending $profile(u_j)$ as a sequence of vectors representing content of the items. α and β are the constants to represent the importance of each score. The procedure for a justified recommendation based on social influence is described below.

We construct a genre precedence information matrix from vector precedence information as shown in Table 1. The matrix VP in Table 1a represents the vector precedence information. The matrix GP in Table 1b indicates the genre precedence information. The entry at VP_{ij} indicates the number of times vector v_i precedes vector v_j. For instance, if we consider a specific instance, say, $[v_2, v_4] = 4$ then it means that v_2 precedes v_4 four times. Similarly if v_2 contains genre g_1 and v_4 contains genre g_2 then the genre precedence information given in Table 1 for $[g_1, g_2]$ should be incremented accordingly. Table 1 shows the simple example of calculating genre precedence information from vector precedence information. For simplicity, we assumed that each item is represented with 5 genres only. Table 4.1(a) shows the vector precedence information. Suppose there are four vectors v_1, v_2, v_3, and v_4 which v_1 contains genres $\{g_1, g_3\}$, v_2 contains the genres $\{g_3, g_5\}$, v_3 contains the genres$\{g_1, g_3\}$ and v_4 contains the genres $\{g_1, g_2$ and $g_4\}$.

To calculate the score we are using *top-I* precedence probabilities for each item, which means I number of consumed items are used from the profile of target user/group(refer Eq. 1). As mentioned earlier, in the case of social influence model we consider *top-I* vectors and *top-I* actors (see Eq. 3). So whatever the score we are getting for each item is because of the *top-I* vectors and *top-I* actors associated with it. Hence, we can claim that these are the justifications for the different scores we are getting. After comparing the scores attached to the different items, we identify the item/s with top score and also determine the *top-I* vectors and actors associated with that item. The information contained in these *top-I* vectors and actors provides reasons/justification for the recommendations. However, a user may not understand if we say that because of some vector V we recommended this for you. So, it is better to identify the corresponding items from the user profile and come with a justification like *because of those items that contain so and so features it is recommended for you*. It would be better if our justification style is in terms of actors, like, *those who watches actor* '*a*' *movie also watches actor* '*b*' *movie*.The clarification behind

this is social impact performs a key role in the recommendation. The style of justification we want to provide looks more or less like "movie x is recommended for you because, those who watches $\underbrace{y, z, \ldots}_{M}$, which contains $\underbrace{\{g_1, g_2, \ldots\}}_{\text{Ģ}}$ also watches x which contain $\underbrace{\{g_1, g_3, \ldots\}}_{\acute{\text{G}}}$ and those who watches actors $\underbrace{a, b, \ldots}_{A}$ movies also watches actor c movies".

To do above kind of justification, we need to know items/movies list M, genre lists Ģ and Ǵ, and actors list A and actor c. What we know is *top-I* vectors, *top-I* actors and recommended movie/item, vector, and actor of a recommended movie. Map the *top-I* vectors to the profile of target user(s)/group and get the corresponding items/movies and this would form the movies list M. To get the genre lists Ģ and Ǵ, find the genre precedence information for every genre g_i present in the *top-I* vectors with every genre g_j present in the vector of a recommended item. For any pair of g_i and g_j, if GP_{ij} exceeds some threshold τ, then add genre g_i to the list Ģ and genre g_j to the list Ǵ. *top-I* actors form a list A and c is simply the actor of a recommended movie.

4.4 Justification in GRS

Panagiotis et.al. [14] introduce keyword based, influence-style justification in Personal Recommender Systems(PRS). Providing *justified recommendation* in GRS is difficult as compared to PRS. The reason being, in GRS we have to satisfy all or most of the group members with our justification or explanation, whereas in PRS, explanation corresponds to a single user and therefore it is easy to satisfy him/her. We analyze different possibilities that group members present geographically in a recommender system scenario.

There are two different ways for the *target group* to logon to the system. (1) All the members logged on one GUI. (2) Every member has a different GUI. When all members logged on single GUI, our justification should be from the group perspective. When every member logged on to multiple GUIs there is no influence factor on each other, so the explanation should be from an individual perspective, though the recommendation should satisfy everyone in the group. Models used for these two different concepts are explained as follows.

Single GUI—When all the group members are at one place, they influence each other [5] and give respect to a majority of the group members interests. From this, it is clear that our goal is to satisfy all or most of the group members while giving justification. We have used *weighted virtual user* for this (Eqs. 1 and 3), where weight ensures percentage of the group watched a particular content/item.

Multiple GUI—When each of the group members is located at a different place, they do not influence each other much [5]. They think from their own perspective, so our justification should be different for different members of the group, i.e., our explanation should cover, best views of individual members that result in the recommendation and we should present those while justifying. We have chosen *merging scores* strategy which in turn aggregates individual scores to get the group score. Give an appropriate explanation for different users in the group, by considering individual *top-I* vectors and actors of an item recommended.

5 Experimental Analysis

We report the experimental results on benchmark dataset, MovieLens 100K,[4] to observe the correctness of the proposed model. We conduct experiments for different group sizes of random selection. All the outcomes disclosed here are normal of 50 runs. To check the correctness of *justification*, we introduced a new measure called *Justification rate*, defined as follows.

Justification rate—Assume that you have recommended an item o_1 for some group G and justifying it by considering *top-I* vectors and actors. Now drop these vectors actors that we are claiming as a reason for the recommendation and recommend again. Say the item recommended in the second iteration as item o_2. If the o_1 and o_2 are not same then it is a *hit*. Count the number of hits over t iterations for different randomly generated groups, for fixed group size. The ratio of the number of hits and t will give the *justification rate*.

$$\textit{Justification rate} = \frac{\sum_{i=1}^{t}(o_1 \neq o_2 \; ? \; 1 : 0)}{t} \tag{4}$$

The idea is, after justifying a recommended item, to remove the vectors and actors from the user profile(s) that we are claiming as the reasons for the recommendation. We recommend again, and if the previously recommended item and item recommended now are not same, then we can say that our justification is correct. To ensure the correctness of our justification we conducted three different experiments.

Experiment-1: Drop every vector and actor that we are claiming as a reason for the recommendation from the target user/group profile then recommend again and check with the previous recommendation.

Our justification style would be very lengthy when *top-I* is high. To overcome this problem we consider including only top vectors/actors in the explanation. We conduct experiments 2 and 3 to analyze the *justification rate* related to this.

[4] https://www.grouplens.org/datasets/movielens/.

Table 2 Justication Rate (JR) of various models for dierent experiments with varying group sizes

Experiments	Model	JR
Experiment-1	Multiple GUI	88
	Single GUI	90
Experiment-2	Multiple GUI	68
	Single GUI	74
Experiment-3	Multiple GUI	68
	Single GUI	64

(a) Group size = 2

Experiments	Model	JR
Experiment-1	Multiple GUI	88
	Single GUI	92
Experiment-2	Multiple GUI	74
	Single GUI	76
Experiment-3	Multiple GUI	70
	Single GUI	68

(b) Group size = 5

Experiments	Model	JR
Experiment-1	Multiple GUI	90
	Single GUI	92
Experiment-2	Multiple GUI	78
	Single GUI	68
Experiment-3	Multiple GUI	82
	Single GUI	50

(c) Group size = 10

Experiments	Model	JR
Experiment-1	Multiple GUI	86
	Single GUI	86
Experiment-2	Multiple GUI	80
	Single GUI	60
Experiment-3	Multiple GUI	78
	Single GUI	38

(d) Group size = 20

Experiment-2: Drop top vector and top actor that we are claiming as a reason for the recommendation from the target user/group profile then recommend again and check with the previous recommendation.
Experiment-3: Drop top vector or top actor that we are claiming as a reason for the recommendation from the target user/group profile then recommend again and check with the previous recommendation.

Table 2 shows the results of three experiments, for different models (same GUI and different GUI), for different group sizes. JR in the table indicates *Justification Rate*. From the tables, it is observed that justification rate is good, which means the justification we are providing is good/perfect for this model. If we feel the explanation is too lengthy, we can just give the content related to top vector/actor only in the justification as the results of Experiment-2 and Experiment-3 are also exhibiting good justification rate.

6 Conclusions

We made an initial attempt to develop a justified group recommendation model in this paper. We generalized the precedence mining model to accomodate social influence factor. A new measure called justification rate has been introduced toward an efficient evaluation of justified recommender systems. Our experimental results are showing

that our justification is valid and correct. In the future, we propose to generalize our framework to accommodate different varieties of group recommender algorithms.

References

1. Amer-Yahia, S., Roy, S.B., Chawla, A., Das, G., Yu, C.: Group recommendation: semantics and efficiency. PVLDB **2**(1), 754–765 (2009)
2. Baltrunas, L., Makcinskas, T., Ricci, F.: Group recommendations with rank aggregation and collaborative filtering. In: Proceedings of the 2010 ACM Conference on Recommender Systems, pp. 119–126 (2010)
3. Bilgic, M.: Explaining recommendations: satisfaction versus promotion. In: Proceedings of Beyond Personalization, the Workshop on the Next Stage of Recommender Systems Research, pp. 13–18 (2005)
4. de Campos, L.M., Fernández-Luna, J.M., Huete, J.F., Rueda-Morales, M.A.: Managing uncertainty in group recommending processes. User Model. User-Adapt. Interact. **19**(3), 207–242 (2009)
5. Jameson, A., Smyth, B.: Recommendation to groups. In: The Adaptive Web, pp. 596–627 (2007)
6. Kagita, V.R., Meka, K.C., Padmanabhan, V.: A novel social-choice strategy for group modeling in recommender systems. In: 2015 International Conference on Information Technology (ICIT), pp. 153–158. IEEE (2015)
7. Kagita, V.R., Padmanabhan, V., Pujari, A.K.: Precedence mining in group recommender systems. In: International Conference on Pattern Recognition and Machine Intelligence, pp. 701–707. Springer (2013)
8. Kagita, V.R., Pujari, A.K., Padmanabhan, V.: Group recommender systems: a virtual user approach based on precedence mining. In: Australasian Joint Conference on Artificial Intelligence, pp. 434–440. Springer (2013)
9. Kagita, V.R., Pujari, A.K., Padmanabhan, V.: Virtual user approach for group recommender systems using precedence relations. Inf. Sci. **294**, 15–30 (2015)
10. Masthoff, J.: Group recommender systems: combining individual models. In: Recommender Systems Handbook, pp. 677–702. Springer (2011)
11. Mobasher, B., Burke, R.D., Bhaumik, R., Sandvig, J.J.: Attacks and remedies in collaborative recommendation. IEEE Intell. Syst. **22**(3), 56–63 (2007)
12. Parameswaran, A.G., Koutrika, G., Bercovitz, B., Molina, H.G.: Recsplorer: recommendation algorithms based on precedence mining. In: SIGMOD, pp. 87–98 (2010)
13. Shin, C., Woo, W.: Socially aware tv program recommender for multiple viewers. IEEE Trans. Consum. Electron. **55**(2), 927–932 (2009)
14. Symeonidis, P., Nanopoulos, A., Manolopoulos, Y.: Providing justifications in recommender systems. IEEE Trans. Syst. Man Cybern. Part A **38**(6), 1–1272 (2008)
15. Yu, Z., Zhou, X., Hao, Y., Gu, J.: Tv program recommendation for multiple viewers based on user profile merging. User Model. User-Adapt. Interact. **16**(1), 6382 (2006)

Part V
Communication Systems, Antenna Research, and Cognitive Radio

Equalization of Communication Channels Using GA-Trained RBF Networks

Pradyumna Mohapatra, Tumbanath Samantara, Siba Prasada Panigrahi and Santanu Kumar Nayak

Abstract Equalization of communication channels still remains a challenge. Radial Basis Function Neural Network (RBFNN) based equalizers are also well known in the literature. However, Design of RBFNN using traditional hit and trial is time-consuming and suboptimal in nature. Hence, this work proposes the optimal design of RBFNN equalizers using Genetic Algorithms (GA). Also, methods used in literature deals equalization problem as an optimization problem. However, this work deals the same as a classification problem. Simulation results prove the better performance of proposed equalizer.

Keywords Equalization · GA · RBFNN

1 Introduction

Communication channel corrupts the information. Recovery of this information is termed as equalization and the corresponding device is the equalizer. The past few decades of literature on equalization have seen wide use of Artificial Neural Networks (ANN) [1–4]. However, RBFNN equalizers deliver better performance [5] because of its advantages like; (a) a compact structure requiring an easy training method and hence consumes lesser time [6] and (b) better ability to generalize ability and also an approximation of nonlinear functions becomes precise [7]. Further, RBFNN delivers a stable equalization with a higher convergence speed [8]. This is further proved in the works of [3, 9–14].

P. Mohapatra (✉) · T. Samantara
Orissa Engineering College, Bhubaneswar, India
e-mail: pradyumnapapers@rediffmail.com

S. P. Panigrahi
VSSUT, Burla 60818, India

S. K. Nayak
Berhampur University, Berhampur, India

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_48

But, traditional methods of designing RBFNN training are hit and trial and also consume more time. GA and PSO have been used for the design of RBFNN respectively in [15] and [16]. In these works, they attempted to finalize the centers, spread, and bias through minimization of MSE between the expected and actual outputs. In this work, we denote GA-trained RBFNN as GRBF.

The major role of this work is the use of GRBF in the equalization of communication channels. The basic difference between the proposed equalizers and those of the methods available in the literature using GA is that this paper treats the problem as a classification problem while those in literature treat the same as an optimization problem, Simulation results performed in this work proves superior performance of proposed equalizer as compared to existing GA-based equalizers [17, 18], and also existing RBF-based equalizers [8, 9].

2 RBFNN Equalizer

Digital communication system considered in this paper is illustrated in Fig. 1. If $x(k)$ is the transmitted data at time instant k, the channel can be modeled like a Finite Impulse Response filter and output, $y_1(k)$, at the same time instant where [5],

$$y_1(k) = \sum_{i=0}^{N-1} h_i x(k-i) \tag{1}$$

Here, $h_i(i = 0, 1, \cdots N-1)$ denotes the tap weights for the channel and N denotes its length. The "NL" block denotes the nonlinear distortion introduced in the. The mostly used nonlinear function is

$$y(k) = F(y_1(k)) = y_1(k) + by_1(k)^3 \tag{2}$$

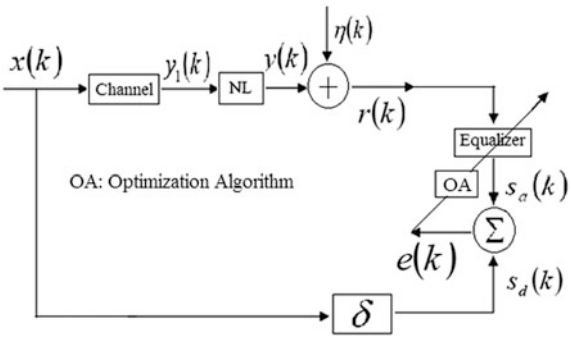


Fig. 1 System model with equalizer

Here b is a constant. Therefore, the channel output becomes

$$y(k) = \left(\sum_{i=0}^{N-1} h_i x(k-i)\right) + b\left(\sum_{i=0}^{N-1} h_i x(k-i)\right)^3 \tag{3}$$

The channel output is once again affected by noise $\eta(k)$. The noise assumed here as additive zero-mean white Gaussian with variance, σ^2. The corresponding output signal that reaches the receiver $r(k)$ is

$$r(k) \cong y(k) + \eta(k) \tag{4}$$

This is the input signal for the equalizer. The job of the equalizer is to recover the original sequence (while considering transmission delay, δ) by nullifying the effects of distortion and noise. $x(k-\delta)$. This signal is termed as the and given by

$$s_d(k) = x(k-\delta) \tag{5}$$

The problem of equalization is a kind of classification problem [2–5], where the job of the equalizer becomes to identify and to bring a dividing wall for different categories in the input $x(k) = [x(k), x(k-1), \cdots x(k-N+1)]^T$ and make two separate regions.

Bay's theory provides an expression for an optimum solution to a nonlinear classification problem by a decision function:

$$f_{bay}(x(k)) = \sum_{j=1}^{N} \beta_j \exp\left(\frac{-\left\|x(k) - c_j\right\|}{2\sigma^2}\right) \tag{6}$$

With binary transmitted symbols:

$$\beta_j = \begin{cases} +1 & for \quad c_j \in C_d^{(+1)} \\ -1 & for \quad c_j \in C_d^{(-1)} \end{cases} \tag{7}$$

Here, $C_d^{(+1)}/C_d^{(-1)}$ is the set of channel states, c_j is a binary symbol, $x(k-\delta) = +1/-1$.

In Fig. 1, the block "Equalizer" denotes RBFNN. GA is used to find the optimal number for layers and neurons in each of these layers (except that for input layer) for this RBFNN and shown by the block "OA". The number of neurons in input layer is same as the number of taps in the channel, N.

The equalizer output is as follows:

$$\begin{aligned} f_{RBF}(s(k)) &= \sum_{j=1}^{z} w_j \exp\left(\frac{-\left\|s(k) - t_j\right\|^2}{\alpha_j}\right) \\ &= W^T(k)\varphi(k) \end{aligned} \tag{8}$$

Here,

$$W(k) = [w_1(k), w_2(k), \cdots w_z(k)]^T$$
$$\varphi(k) = [\phi_1(k), \phi_2(k), \cdots \phi_z(k)]^T$$
$$\phi_j = \exp\left(\frac{-\|s(k) - t_j\|^2}{\alpha_j}\right) \quad \text{for } j = 1, 2, \cdots z$$

Here, t_j and α_j respectively, denotes the centers and the spreads in hidden layers and w_j denotes the connecting weights.

Equation (8) that is an implementation of the Bay's decision function of Eq. (6) considers t_j same as the channel states, c_j with the adequately regulated connecting weights.

Therefore, the decision function for the RBFNN equalizer is

$$s_d(k-\delta) = \begin{cases} +1 & f_{RBF}(x(k)) \geq 0 \\ -1 & otherwise \end{cases} \tag{9}$$

The divergence of the RBFNN equalizer output $(\hat{x}(k-\delta))$ from the expected output $(x(k-\delta))$ is termed as the error and denoted by $e(k)$. This error is used in updation of the weights of the equalizer. Two popular indexes for performance are, MSE and BER, $E[e(k)]$. Here, E is the expectation operator.

3 GA-Trained RBFNN Equalizer, GRBF

GA is one of the most popular methods for optimization and search problems [19, 20]. In this paper, GA-based equalizers are developed through selection of a set within a bound $\{\pm 1\}$ as chromosomes representing the weight vector for the equalizer. GA starting with a random initial string and using 3 operators repeatedly flows to iterations to come. Individuals with the best fitness are given by the selection and provide minimum MSE at the equalizer output.

As mentioned earlier, Barreto et al. in [15] used GA used PSO for the design of RBFNN. Since the RBFNN has only one hidden layer, use of GA for architecture optimization becomes simpler. Also, upper bound on the number of neurons in the hidden layer is decided by the number of data set during the training. Optimal determination of the number of basis centers is achieved using GA. In this work, GA is used to optimize the architecture of RBFNN and the next stage uses supervised learning. The distance between the input vector and the centers is denoted as distance factor.

Steps used in the training used in this work are as follows:

i. Randomly initialize the distance factor;
ii. Design the RBFNN architecture (define the number of centroids, decided by GA).
iii. Update weights of output layer
iv. Evaluate fitness using GA
v. If stopping criteria met, then stop training,
Else, go to step i.

Use of GA in the design of RBFNN, termed here as GRBF, has been reviewed in [21]. Use of GRBF in Classification [22], function approximation [23], time series prediction [24], etc., also helps this work for use in channel equalization.

4 Simulations

For evaluation of performance, we have considered two channels with a transfer function as follows:

$$H_1(z) = [0.26 \quad 0.93 \quad 0.26] \tag{10}$$

$$H_2(z) = [1 \quad 1 \quad 1] \tag{11}$$

Nonlinearity and noise as discussed in Eq. (2) are introduced. Noise variance was chosen as unity. As discussed earlier, MSE and BER were chosen as a performance index. To estimate MSE, a constant Signal-to-Noise (SNR) of 10 dB is chosen. Simulations were conducted for 300 iterations and averaged.

Simulation parameters for GA used are same as [25] and outlined in Table 1. For performance comparisons, we have compared the proposed equalizer with GA-U [21], GA-CS [23], RBF [9], and MCRBF [10] equalizers.

Performance of the equalizers for the channels of Eqs. (10) and (11) are, respectively, illustrated through their MSE as in Figs. 2 and 3. Evaluation of Fig. 2 for the channel of Eq. (10), we can summarize in the following manner:

Table 1 Simulation parameters for GA

Parameter	Value	Parameter	Value
Selection type	Roulette	Mutation type	Uniform
Mutation ratio	0.03	Crossover type	Singe point
Cross over ratio	0.9	Population size	50

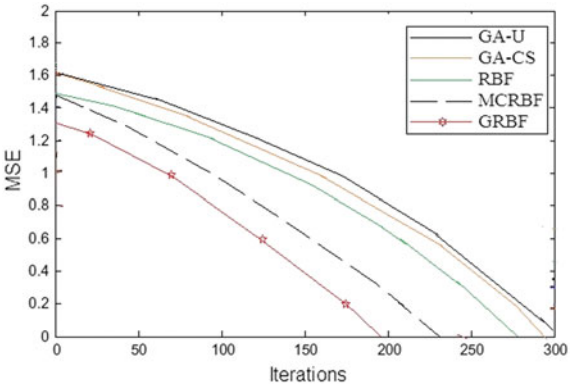


Fig. 2 MSE for channel of Eq. (10)

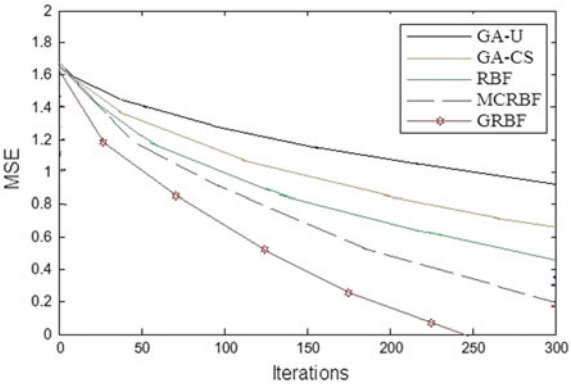


Fig. 3 MSE for channel of Eq. (11)

- It is observed that GA-CS, RBF, MCRBF, and proposed GRBF equalizers converges after 294, 278, 228, and 196 iterations respectively. However, GA-U failed to converge even within 300 iterations.
- GA-U and GA-CS performances are almost similar and comparable to that of RBF after 50 iterations, with RBF performance slightly better than the other two.
- There is 0.2 dB of difference exists between MSE of MCRBF and GRBF at each iteration with GRBF performing better than MCRBF.

Evaluation of Fig. 3 for channel of Eq. (11), we can summarize in the following manner:

- It is observed that GRBF equalizers converge after 248 iterations. However, all other equalizers failed to converge within 300 iterations.

Performance of the equalizers for the channels of Eqs. (10) and (11) are also studied through their BER as in Figs. 4 and 5. Evaluation of Fig. 4 for channel of Eq. (10), we can summarize in the following manner:

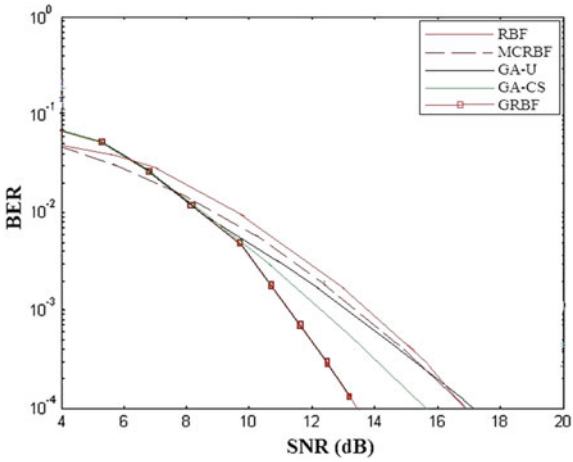


Fig. 4 BER for channel of Eq. (10)

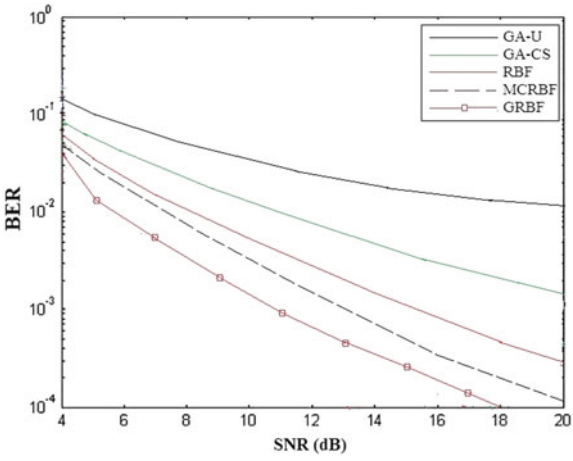


Fig. 5 BER for channel of Eq. (11)

- Performance of all equalizers is comparable to each other up to SNR of 9 dB.
- After SNR of 9 dB, GRBF outperforms other equalizers

Evaluation of Fig. 3 for the channel of Eq. (11), we can summarize in the following manner:

- For all SNR values, GRBF outperforms other equalizers.

5 Conclusion

A novel GRBF-based equalizer is proposed in this paper. Here GA trains RBFNN equalizer. GRBF equalizer outperforms contemporary GA- and RBF-based equalizers at all conditions of noise. The proposed equalizer found to be robust, stable and efficient because of inherent advantages of RBFNN. This theoretical background of better performance may provoke the research in the area to reach the milestones and may appear in future works.

References

1. Seyman, M.N., Taşpınar, N.: Channel estimation based on neural network in space time block coded MIMO–OFDM system. Digit. Signal Proc. **23**, 275–280 (2013)
2. Ruan, X., Zhang, Y.: Blind sequence estimation of MPSK signals using dynamically driven recurrent neural networks. Neurocomputing **129**, 421–427 (2014)
3. Rizaner, A.: Radial basis function network assisted single-user channel estimation by using a linear minimum mean square error detector under impulsive noise. Comput. Electr. Eng. **39**, 1288–1299 (2013)
4. Sahoo, H.K., Dash, P.K., Rath, N.P.: NARX model based nonlinear dynamic system identification using low complexity neural networks and robust H_∞ filter. Appl. Soft Comput. **13**, 3324–3334 (2013)
5. Chen, S., Mulgrew, B., Grant, M.P.: A clustering technique for digital communication channel equalisation using radial basis function network. IEEE Trans. Neural Netw. **4**, 570–579 (1993)
6. Qasem, S.N., Shamsuddin, A.M., Zain, A.M.: Multi-objective hybrid evolutionary algorithms for radial basis function neural network design. Knowl. Based Syst. **27**, 475–497 (2012)
7. Dong, X., Wang, C., Zhang, Z.: RBF neural network control system optimized by particle swarm optimization. In: 3rd IEEE International Conference on Computer Science and Information Technology, pp. 348–351. Chengdu-China (2010)
7. Kahphooi, S., Zhihong, M., Wu, H.R.: Nonlinear adaptive RBF neural filter with Lyapunov adaptation algorithm and its application to nonlinear channel equalization. In: Proceedings of the Fifth International Symposium on Signal Processing and its Applications, vol. 1, pp. 151–154 (1999)
9. Rajbhandari, S., Faith, J., Ghassemlooy, Z., Angelova, M.: Comparative study of classifiers to mitigate intersymbol interference in diffuse indoor optical wireless communication links. Optik Int. J. Light Electron Optics **124**, 4192–4196 (2013)
10. Zeng, X., Zhao, H., Jin, W., He, Z., Li, T.: Identification of nonlinear dynamic systems using convex combinations of multiple adaptive radius basis function networks. Measurement **46**, 628–638 (2013)
11. Yee, M.S., Yeap, B.L., Hanzo, L.: Radial basis function-assisted turbo equalization. IEEE Trans. Commun. **51**, 664–675 (2003)
12. Xie, N., Leung, H.: Blind equalization using a predictive radial basis function neural network. Trans. Neural Netw. **16**, 709–720 (2005)
13. Lee, J., Sankar, R.: Theoretical derivation of minimum mean square error of RBF based equalizer. Sig. Process. **87**, 1613–1625 (2007)
14. Li, M.N., Huang, G.B., Saratchandran, P., Sundarrajan, N.: Performance evaluation of GAP-RBF network in channel equalization. Neural Process. Lett. **22**, 223–233 (2005)

15. Barreto, A.M.S., Barbosa, H.J.C., Ebecken, N.F.F.: Growing compact RBF networks using a genetic algorithm. In: Proceedings of the VII Brazilian Symposium on Neural Networks, pp. 61–66 (2002)
16. Feng, H.M.: Self-generating RBFNs using evolutional PSO learning. Neurocomputing **70**, 241–251 (2006)
17. Nazmat, S., Zhu, X., Gao, J.: Genetic algorithm based equalizer for ultra-wideband wireless communication systems, chapter 4. In: Ultra Wideband Communications: Novel Trends—System, Architecture and Implementation (2011). ISBN 978–953-307-461-0
18. Altın, G., Martin, R.K.: Bit-error-rate-minimizing channel shortening using post-FEQ diversity combining and a genetic algorithm. Sig. Process. **91**, 1021–1031 (2011)
19. Mandal, A., Zafar, H., Das, S., Vasilakos, A.V.: Efficient circular array synthesis with a memetic differential evolution algorithm. Prog. Electromagnet. Res. **38**, 367–385 (2012)
20. Goldberg, D.E.: Genetic Algorithm in Search Optimization and Machine Learning. Addison Wesley (1989)
21. Harpham, C., Dawson, C.W., Brown, M.R.: A review of genetic algorithms applied to training radial basis function networks. Neural Comput. Appl. **13**, 193–201 (2004)
22. Kurban, T., Besdok, E.: A comparison of RBF neural network training algorithms for inertial sensor based terrain classification. Sensors **9**, 6312–6329 (2009)
23. Awad, M.: Optimization RBFNNs parameters using genetic algorithms: applied on function approximation. Int. J. Comput. Sci. Secur. **4**, 295–307 (2010)
24. Gan, M., Pend, X., Dong, X.: A hybrid algorithm to optimize RBF network architecture and parameters for nonlinear time series prediction. Appl. Math. Modell. **36**(7), 2911–2919 (2012)
25. Mohammed, J.R.: A study on the suitability of genetic algorithm for adaptive channel equalization. Int. J. Electr. Comput. Eng. **2**, 285–292 (2012)

Effect of Circular Variation in Thickness and Linear Variation in Density on Vibrational Frequencies

Amit Sharma, Ashok Kumar Raghav, Vijay Kumar and Ashish Kumar Sharma

Abstract Plates with variable thickness are widely used because of their high strength and durability. An effect of circular variation in thickness along with linear variation in density on thermal-induced vibration is studied. The geometry of the plate is considered to be rectangular. The thickness, as well as density variation of the plate, is to be taken one dimensional. For thermal effect, the temperature of plate is to be taken bi-parabolic. To solve the governing differential equation, Rayleigh–Ritz technique is applied for different value of nonhomogeneity, thermal gradient, tapering parameter, and aspect ratio. Results are calculated for Duralium material with high-level computational software MAPLE and presented with the help of graphs.

Keywords Thermal gradient · Parabolic variation · Circular variation Vibration

1 Introduction

In the modern world, plates of different shape and size with nonuniform thickness and temperature are widely used in various engineering and mechanical structures such as a nuclear reactor, missiles, and wings of an aircraft. By using appropriate variation in thickness, nonuniformity in plates arises which have considerably better

A. Sharma (✉) · A. K. Raghav · V. Kumar
Amity University Haryana, Gurgaon 122413, India
e-mail: dba.amitsharma@gmail.com

A. K. Raghav
e-mail: akraghava@gmail.com

V. Kumar
e-mail: vkb1605@gmail.com

A. K. Sharma
Surya World Group of Institution, Rajpura, Punjab, India
e-mail: ashishk482@gmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_49

competency than the uniform plates. The consideration of nonuniformity along with nonhomogeneity in plate's material not only ensures the high strength in engineering situation but also fulfill the requirements of the industry of aerospace, ocean engineering, and optical equipment. Therefore, for making such structures it is necessary to study plate characteristic during vibration under nonuniformity, nonhomogeneity and temperature fields. The first few modes of vibration play a tremendous role to make more authenticate and reliable structures. In the available literature, a large number of works has been done by taking either a linear variation or parabolic variation in thickness under thermal gradient but none of them studied the effect circular variation in thickness.

Transverse vibration of the simply supported plate with the oblique cut is studied by Avalos and Laura [1]. An exact solution of an eccentric elliptical plate has been discussed by Hasheminejad et al. [2]. The vibration of a rectangular plate with linear thickness has been studied by Gupta and Khanna [3]. Gupta et al. [4] analyzed temperature effects on nonuniform and nonhomogeneous orthotropic rectangular plate. Tomar and Gupta [5, 6] discussed thermal effects on frequencies on nonuniform orthotropic rectangular plate. Sharma et al. [7] analyzed vibrations of an orthotropic rectangular plate and find deflection function, time period, etc. An effect of Poisson's ratio and temperature on vibrational modes using rectangular plate has been discussed by Khanna et al. [8]. Khanna and Kaur [9–12] studied the vibration on a rectangular plate by taking different variation in parameters. They studied effect of bilinear and bi-parabolic temperature with linear thickness variation with varying Poisson ratio as nonhomogeneity. Dhotarad et al. [13] studied vibrational analysis of plate under temperature effect. Leissa [14] studied vibration of the plate and provide a monograph which contains different combinations of variation of parameters. Patel et al. [15] presented influence of stiffener on vibrational behavior of rectangular plate with simply supported edges.

In the present problem, authors studied the effect of circular variation in thickness and linear variation in density along with two-dimensional temperatures on the vibration of rectangle plate. Rayleigh–Ritz technique is applied to find vibrational frequencies for various components such as temperature, aspect ratio, nonhomogeneous parameter and tapering constants. Results are displayed with the help of graphs.

2 Governing Equation of Motion

For the plate having rectangular geometry, the mathematical expression for differential equation is

$$\left[\begin{array}{l} D_1\left(\Phi_{,\zeta\zeta\zeta\zeta}+2\Phi_{,\zeta\zeta\psi\psi}+\Phi_{,\psi\psi\psi\psi}\right)+2D_{1,\zeta}\left(\Phi_{,\zeta\zeta\zeta}\Phi_{,\zeta\psi\psi}\right)+2D_{1,\psi}\left(\Phi_{,\psi\psi\psi}+\Phi_{,\psi\zeta\zeta}\right) \\ +D_{1,\zeta\zeta}\left(\Phi_{,\zeta\zeta}+\nu\Phi_{,\psi\psi}\right)+D_{1,\psi\psi}\left(\Phi_{,\psi\psi}+\nu\,\Phi_{,\zeta\zeta}\right)+2(1-\nu)D_{1,\zeta\psi}\Phi_{,\zeta\psi} \end{array}\right] -\rho p^2 T\Phi=0 \tag{1}$$

where D_1, T, ρ, Φ and ν is known as flexural rigidity, the thickness of the plate, density, two-term deflection function, and Poisson's ratio of the rectangular plate. A comma in Eq. (1) is known as partial derivative of Φ with respect to associated independent variable.

The expression for flexural rigidity is

$$D_1 = YT^3/12\left(1-\nu^2\right) \tag{2}$$

where Y is known as Young's modulus.

The expression of Young's modulus with temperature dependence is taken as

$$Y = Y_0(1-\gamma\tau) \tag{3}$$

τ, Y_0 represents temperature on the plate, Young's modulus at the temperature τ_0 and γ is known as the slope of variation.

It is also taken into account that temperature of the plate is in bi-parabolic variation, therefore

$$\tau = \tau_0\left(1-\frac{\zeta^2}{a^2}\right)\left(1-\frac{\psi^2}{b^2}\right) \tag{4}$$

Using Eq. (4), Eq. (3) becomes

$$Y = Y_0\left(1-\alpha\left(1-\frac{\zeta^2}{a^2}\right)\left(1-\frac{\psi^2}{b^2}\right)\right) \tag{5}$$

and $\alpha = \gamma\tau_0 (0 \le \alpha < 1)$ is known as temperature gradient.

Also, the plate's thickness is considered to be circular in one dimension as

$$T = T_0\left(1+\beta\sqrt{1-\frac{\zeta^2}{a^2}}\right) \tag{6}$$

where $\beta (0 \le \beta \le 1)$ is known as taper constant and thickness become constant, i.e., $T = T_0$ at $\beta = 0$.

The material of the plate is taken as nonhomogeneous, therefore linear variation in density is considered as

$$\rho = \rho_0 \left(1 + \alpha_1 \frac{\zeta}{a}\right) \tag{7}$$

where $\alpha_1 (0 \le \alpha_1 \le 1)$ represents nonhomogeneity constant and $\rho = \rho_0$ at $\zeta = 0$.

On putting the value of Young's modulus and thickness of the plate from Eqs. (5) and (6), Eq. (2) becomes

$$D_1 = \frac{Y_0 T_0^3 \left(1 - \alpha\left(1 - \frac{\zeta^2}{a^2}\right)\left(1 - \frac{\psi^2}{b^2}\right)\right)\left(1 + \beta\sqrt{1 - \frac{\zeta^2}{a^2}}\right)^3}{12(1 - \nu^2)} \tag{8}$$

The limiting conditions for the plate is

$$\begin{matrix} \Phi(\zeta, \psi) = \Phi_{,\zeta}(\zeta, \psi) = 0 \\ \Phi(\zeta, \psi) = \Phi_{,\psi}(\zeta, \psi) = 0 \end{matrix} \text{at} \begin{matrix} \zeta = 0, a \\ \psi = 0, b \end{matrix} \tag{9}$$

Also, the deflection function satisfying Eq. (9) given by [14] as

$$\Phi(\zeta, \psi) = \Omega_1 \left(\frac{\zeta}{a}\right)^2 \left(\frac{\psi}{b}\right)^2 \left(1 - \frac{\zeta}{a}\right)^2 \left(1 - \frac{\psi}{b}\right)^2 + \Omega_2 \left(\frac{\zeta}{a}\right)^3 \left(\frac{\psi}{b}\right)^3 \left(1 - \frac{\zeta}{a}\right)^3 \left(1 - \frac{\psi}{b}\right)^3 \tag{10}$$

where Ω_1 and Ω_2 are known as constants.

3 Solution of the Problem

To solve the problem, Rayleigh–Ritz technique is applied. This method based upon the principle that the maximum potential energy or strain energy V_s is equal to maximum kinetic energy T_s. Therefore,

$$\delta(V_s - T_s) = 0 \tag{11}$$

Where

$$V_s = \frac{1}{2} \int_0^a \int_0^b D_1 \left[\left(\Phi_{,\zeta\zeta}\right)^2 + \left(\Phi_{,\psi\psi}\right)^2 + 2\nu \, \Phi_{,\zeta\zeta} \Phi_{,\psi\psi} + 2(1 - \nu)\left(\Phi_{,\zeta\psi}\right)^2 \right] d\psi d\zeta \tag{12}$$

and

$$T_s = \frac{1}{2}p^2 \int_0^a \int_0^b \rho T \Phi^2 d\psi d\zeta \tag{13}$$

Now, taking nondimensional variable as

$$\zeta_1 = \frac{\zeta}{a}, \psi_1 = \frac{\psi}{a} \tag{14}$$

From Eq. (14), Eqs (12), and (13), converted into

$$V_s^* = L \int_0^1 \int_0^{b/a} \left\{ \left[\frac{\left(1-\alpha\left(1-\zeta_1^2\right)\left(1-\psi_1^2\right)\right)}{\left(1+\beta\sqrt{1-\zeta_1^2}\right)^3} \right] \left[\begin{array}{l} \left(\Phi_{,\zeta_1\zeta_1}\right)^2 + \left(\Phi_{,\psi_1\psi_1}\right)^2 + 2\nu\Phi_{,\zeta_1\zeta_1}\Phi_{,\psi_1\psi_1} \\ +2(1-\nu)\left(\Phi_{,\zeta_1\psi_1}\right)^2 \end{array} \right] \right\} d\psi_1 d\zeta_1 \tag{15}$$

$$T_s^* = \frac{1}{2}\rho_0 p^2 a^2 T_0 \int_0^1 \int_0^{b/a} \left(1+\alpha_1\zeta_1\right)\left(1+\beta\sqrt{1-\zeta_1^2}\right)\Phi^2 d\psi_1 d\zeta_1 \tag{16}$$

Where $L = \frac{Y_0 T_0^3}{24a^2(1-\nu^2)}$

By using Eqs. (15) and (16), Eq. (11), becomes

$$\delta\left(V_s^* - \lambda^2 T_s^*\right) = 0 \tag{17}$$

$\lambda^2 = \frac{12a^4p^2\rho_0\left(1-\nu^2\right)}{Y_0 T_0^2}$ represents parameters of frequencies. Equation (17) consists of two unknowns Ω_1 and Ω_2 because of deflection function. These two unknowns can be obtained by

$$\frac{\partial}{\partial\Omega_n}\left(V_s^* - \lambda^2 T_s^*\right) = 0, \quad n = 1, 2 \tag{18}$$

On simplification of Eq. (18), we get

$$c_{n1}\Omega_1 + c_{n2}\Omega_2 = 0, \qquad n = 1, 2 \tag{19}$$

where c_{n1} and c_{n2} comprise the parametric constant and frequency λ^2

To obtained frequency equation, the determinant of matrix obtained by Eq. (19) must be zero, therefore we get,

$$\begin{vmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{vmatrix} = 0 \tag{20}$$

Equation (20) is a quadratic equation which gives two modes of vibration λ_1 and λ_2

4 Result and Discussion

A numerical illustration is done for Duralium material for finding the frequencies of the present problem by using Eq. (20) corresponding to different aspect's such as a tapering parameter, constant of nonhomogeneity, temperature, and aspect ratio. For such calculation, following values are taken into consideration and results are displayed in the form of graphs.

$$Y_0 = 7.07 * 10^{10}\ \mathrm{N/M^2}, \rho_0 = 2.80 * 10^3\ \mathrm{Kg/M^3}, \nu = 0.345,\ \&\ T_0 = 0.01\ \mathrm{m}$$

Figure 1 shows vibrational frequencies modes for the following three cases

$$\text{Case 1}\ \alpha_1 = \beta = 0, a/b = 1.5,\ \text{Case 2}\ \alpha_1 = \beta = 0.2, a/b = 1.5,$$
$$\text{Case 3}\ \alpha_1 = \beta = 0.6, a/b = 1.5$$

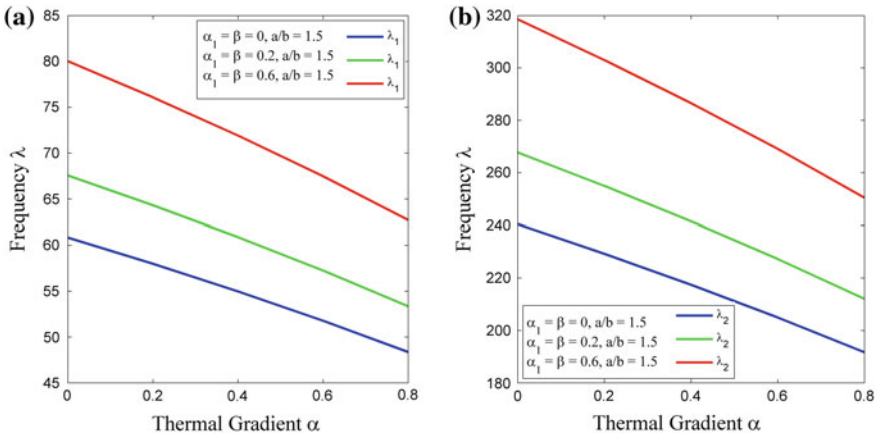


Fig. 1 Thermal gradient α versus frequency λ

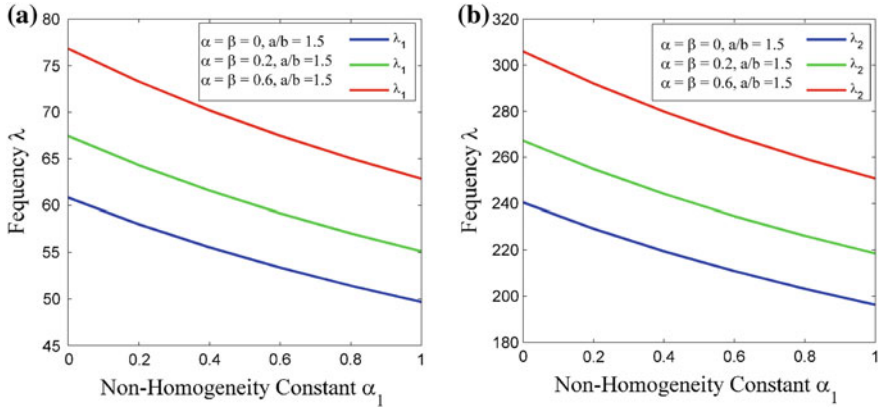


Fig. 2 Nonhomogeneity constant α_1 versus frequency λ

It is concluding that frequency modes decrease with the increase in temperature α from 0 to 0.8. It is also evident that frequency increases when nonhomogeneity α_1 and taper constant β increases from Case 1 to Case 3.

Figure 2 sketches both modes of a vibrational frequency corresponding to nonhomogeneity constant for the following three cases.

$$\text{Case } 4\ \alpha = \beta = 0, a/b = 1.5,\ \text{Case } 5\ \alpha = \beta = 0.2, a/b = 1.5,$$
$$\text{Case } 6\ \alpha = \beta = 0.6, a/b = 1.5$$

One can easily conclude that with the increment in value of nonhomogeneity parameter α_1 from 0 to 1, frequency of vibrational modes decreases. On the other hand frequency increases while moving from Case 4 to Case 6 for the combination of the value of thermal gradient α and taper constant β.

Figure 3 sketches both vibrational modes for fixed temperature gradient $\alpha = 0.4$ corresponding to taper constant β for the subsequent three cases.

$$\text{Case } 7\ \alpha_1 = 0, a/b = 1.5,\ \text{Case } 8\ \alpha_1 = 0.2, a/b = 1.5,$$
$$\text{Case } 9\ \alpha_1 = 0.6, a/b = 1.5$$

From Fig. 3, it is evident that frequency of vibration increases with the increase in taper constant β, while on moving from Case 7 to Case 9, frequency parameter λ decreases.

Figure 4 shows both modes of a vibrational frequency corresponding to the aspect ratio a/b for constant value of thermal gradient $\alpha = 0.4$ and taper constant $\beta = 0.4$. Figure 4 consists of three different cases as

$$\text{Case } 10\ \alpha_1 = 0,\ \text{Case } 11\ \alpha_1 = 0.2,\ \text{Case } 12\ \alpha_1 = 0.6$$

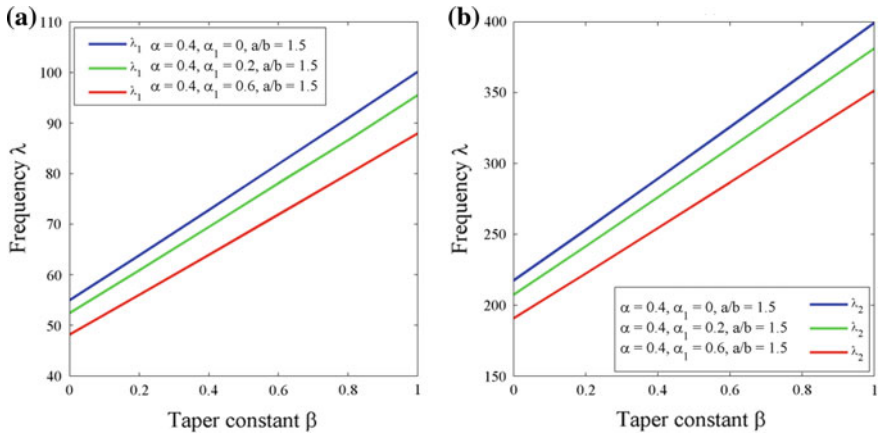


Fig. 3 Taper constat β versus frequency λ

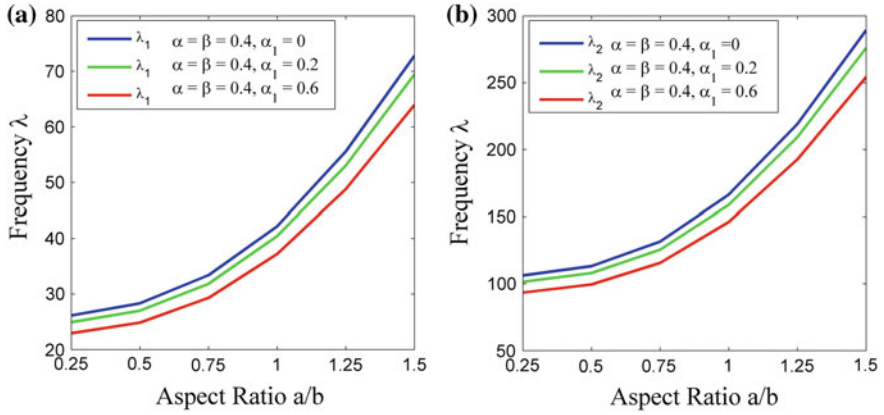


Fig. 4 Aspect ratio (a/b) versus frequency λ

From Fig. 4, one can conclude that frequencies increases when aspect ratio a/b increases from 0.25 to 1.5. But on the other hand, frequencies decreases when nonhomogeneity constant α_1 increases from Case 10 to Case 12.

5 Conclusion

From the above discussion, the author concludes that variation in specific parameter affects the frequency of vibration. Also, the frequency obtained in the present paper (circular variation in thickness and linear variation in density) is less than the

frequency obtained by [9] (linear variation in thickness and exponential variation in Poisson's ratio).

Therefore by choosing an appropriate variation in the parameter or depending upon the requirement, the required frequency can be attained.

References

1. Avalos, D.R., Laura, P.A.: Transverse vibration of simply supported plate of generalized anisotropy with an oblique cut-out. J. Sound Vib. **258**(2), 773–776 (2002)
2. Hasheminejad, S.M., Ghaheri, A., Fadaee, M.: Exact solution for free in plane vibration analysis of an eccentric elliptical plate. Acta Mech. **224**(8), 1609–1624 (2013)
3. Gupta, A.K., Khanna, A.: Vibration of Visco-Elastic Rectangular Plate with Linear Thickness Variation in Both Directions. J. Sound Vib. **301**(3–4), 450–457 (2007)
4. Gupta, A.K., Jhori, T., Vats, R.P.: Study of thermal gradient effect on vibration of a non-homogeneous orthotropic rectangular plate having bi-direction linearly thickness variations. Meccanica **45**(3), 393–400 (2010)
5. Tomar, J.S., Gupta, A.K.: Thermal effect of frequencies of an orthotropic rectangular plate of linearly varying thickness. J. Sound Vib. **90**(3), 325–331 (1983)
6. Tomar, J.S., Gupta, A.K.: Effect of thermal gradient on frequencies of an orthotropic rectangular plate whose thickness varies in two directions. J. Sound Vib. **98**(2), 257–262 (1985)
7. Sharma, A., Sharma, A.K., Ragav, A.K., Kumar, V.: Effect of vibration on orthotropic visco-elastic rectangular plate with two dimensional temperature and thickness variation. Indian J. Sci. Technol. **9**(2), 7 (2016)
8. Khanna, A., Kaur, N., Sharma, A.K.: Effect of varying poisson ratio on thermally induced vibration of non-homogeneous rectangular plate. Indian J. Sci. Technol. **5**(9), 3263–3267 (2012)
9. Khanna, A., Kaur, N.: Effect of thermal gradient on natural frequencies of tapered rectangular plate. Int. J. Math. Anal. **7**, 755–761 (2013)
10. Khanna, A., Kaur, N.: Vibration of non-homogeneous plate subjected to thermal gradient. J. Low Freq. Noise Vib. Active Control **33**(1), 13–26 (2013)
11. Khanna, A., Kaur, N.: A study on vibration of tapered rectangular plate under non-uniform temperature field. Mechanika **20**(4), 376–381 (2014)
12. Khanna, A., Kaur, N.: Effect of structural parameters on the vibrational response of a visco-elastic rectangular plate with clamped ends. Proc. Est. Acad. Sci. **64**(2), 127–138 (2015)
13. Dhotarad, M.S., Ganesan, N.: Vibration analysis of rectangular plate subjected to a thermal gradient. J. Sound Vib. **60**(4), 481–497 (1978)
14. Leissa, A.W.: Vibration of plates (NASA SP-160) (1969)
15. Patel, D.S., Pathan, S.S., Bhoraniya, I.H.: Influence of stiffeners on the natural frequencies of rectangular plate with simply supported edges. Int. J. Eng. Res. Technol. **1**(3), 1–6 (2012)

Design of a Low-Power ALU and Synchronous Counter Using Clock Gating Technique

Nehru Kandasamy, Nagarjuna Telagam and Chinthada Devisupraja

Abstract The need of clock gating for ALU and the synchronous counter is important for the memory system. The synchronous counter using D flip-flop is analyzed with and without clock gating approach and further, it is applied to eight transistors based low-power ALU. The enabled clock-based power gating is considered for ALU and counter circuits. The enabled clock-based gating technique minimizes the low-power comparisons to existing clock gating techniques and results are simulated in tanner software with 130 μm technology.

Keywords CMOS · Clock gating · ALU · Counter · Multiplexer Flip-flop

1 Introduction

The power gating structure is used to eliminate the intermediate nodes for removing leakage current and spurious switching activity for circuits operating on off time. This method is used to maximize the current flow and few n-MOS and p-MOS transistor are used as sleep transistors were evaluated in [1]. The sleep transistor is used to avoid leakage current and maintaining the circuit in low power consumption. In 45 nm technology mode, the leakage current consumes the most of the chip power. In paper [2], researchers have analyzed various leakage current reduction methods. The effective sleep transistors need the optimization of the aspect ratio of transistors and body bias. The address of the different methods of clock gating approaches are register-level clock gating, transparent clock gated pipelines and

N. Kandasamy (✉) · N. Telagam · C. Devisupraja
Institute of Aeronautical Engineering, Hyderabad, India
e-mail: nnehruk@gmail.com

N. Telagam
e-mail: nagarjuna473@gmail.com

C. Devisupraja
e-mail: chinthada.devisupraja@gmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_50

elastic pipeline clock gating for high-end microprocessors. In order to increase the efficiency of the clock gating technique is done by combining several techniques and applies to microprocessor units for further reduction in the leakage power reported in [3].

The paper [4, 5] authors have discussed the issues of surge current properties in distributed network structure with TSMC 90-nm technology. During on the time of the transistor, surge current leads into failure in a clock gating technique. The unwanted switching activity is the major occurrence in total energy consumption. The wake-up schedule is used for subthreshold current. The evaluating importance's of power gating design in microscale for VLSI sub-systems and identifies the values of threshold voltage [6]. The sleep transistor is used to save power in most of the systems. The two important leakage current types are subthreshold leakage and tunneling leakage [7].

On paper [8, 9] researcher deals with the properties and sizing of distributed sleep transistor network using clustered gates for minimizing switching current in functional units. The issues of sleep transistor insertion and sizing problems were discussed in [10]. The number of switching transitions depends on activation period of the clock signal [11].The researchers [12, 13] demonstrated the issues of the mixed integer linear program is analyzed for a wake-up scheduling technique to minimize the power consumption in the function units of the system.

2 Synchronous Counter and Arithmetic Logic Unit

In existing method, the synchronous counter is designed using D flip-flop and clock signal uses more energy. The suggested method for clock gating technique is implemented instead of clock signals; this consumes minimum energy compared to existing methods.

In this paper, the synchronous counter is used to design multiplexer. The counter output is used instead of multiplexer select signals. The proposed multiplexer design is used for arithmetic and logic unit.

2.1 Synchronous Counter with Clock Gating

In Fig. 1 shows the synchronous counter using clock gating technique. The energy consumption can be controlled by using the clock net in gated technique. In memory devices, the timing signal is responsible for a considerable amount of energy dissipation. The gating effect minimizes the switching activity by controlling the clock signal.

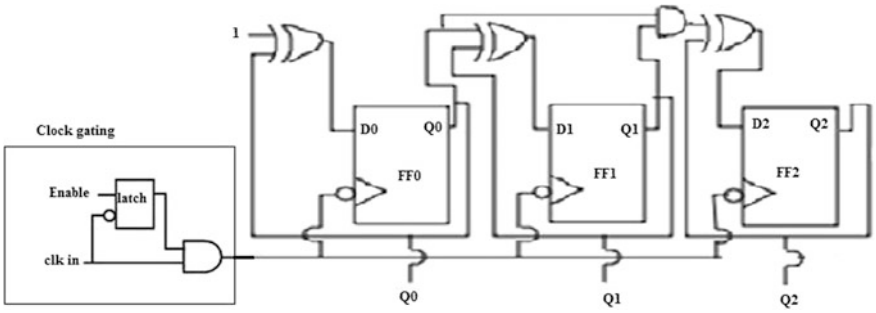
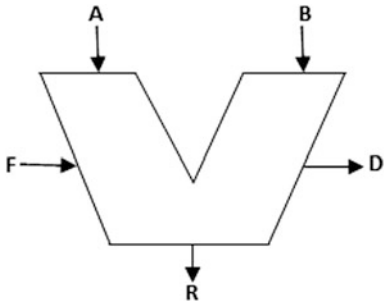


Fig. 1 Synchronous counter using enabled clock gating

Fig. 2 Operands involved in ALU



2.2 Arithmetic and Logic Unit

It plays an important role of designing finite impulse response filter for computing arithmetic and logical operations. Figure 2 shows the representation of the ALU.

ALU is implemented using data selectors and arithmetic units. In full adder, circuit was designed by using eight transistors-based adder circuits for low power consumption.

The output of the synchronous counter is given as the select signal input to the multiplexes. In Fig. 3 shows the general block diagram of the ALU. The proposed circuit is operated based on the control inputs of the data selector [14].

3 Results and Discussions

3.1 Synchronous Counter

The schematic view of the synchronous counter is shown in Fig. 5. The synchronous counter is designed using a group of delay flip-flops and it is shown in Fig. 4.

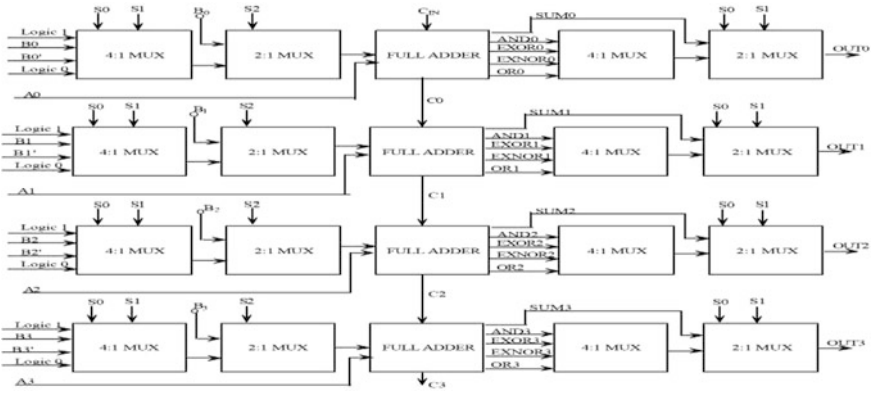


Fig. 3 4-bit ALU using data selectors and adders

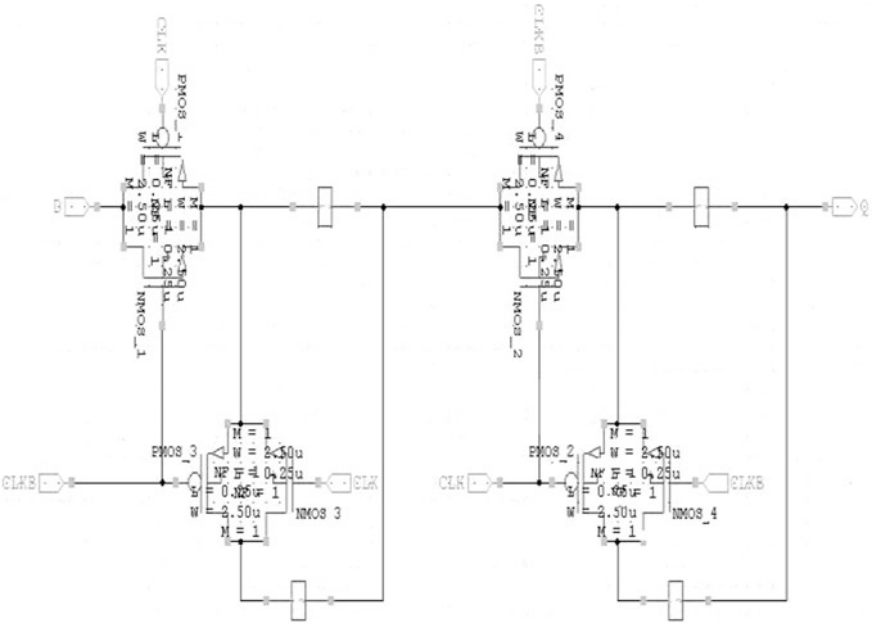


Fig. 4 Schematic diagram of D flip-flop using S-edit

The synchronous counter using delay flip-flop with enabled clock gating technique is shown in Fig. 6.

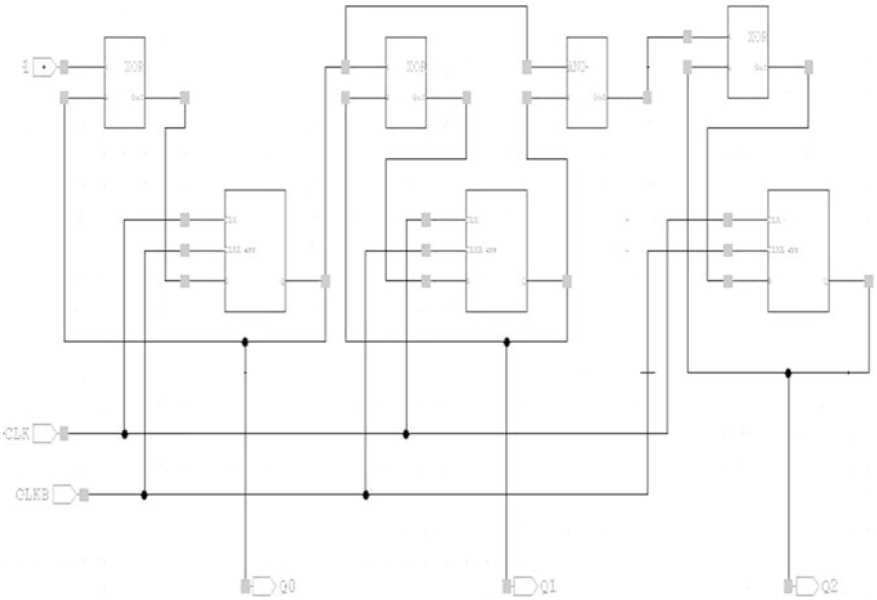


Fig. 5 Synchronous counters using D flip-flop without clock gating technique

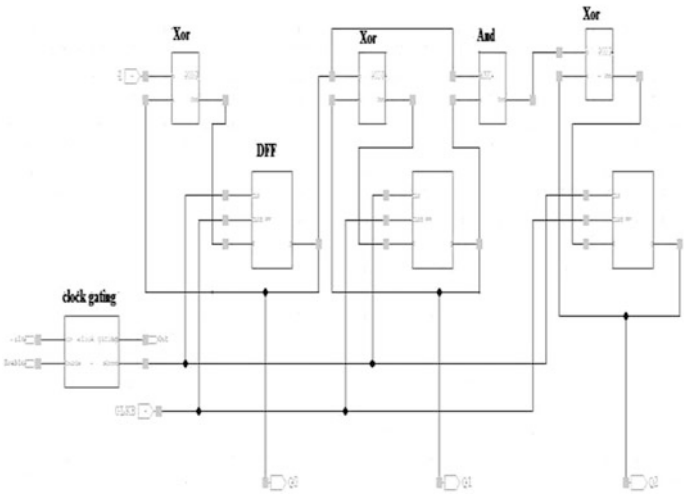


Fig. 6 Synchronous counters using D flip-flop with clock gating technique

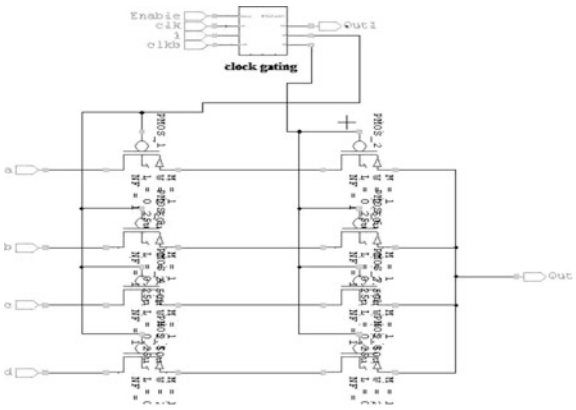


Fig. 7 Design of 4 × 1 data selector using counters with clock gating

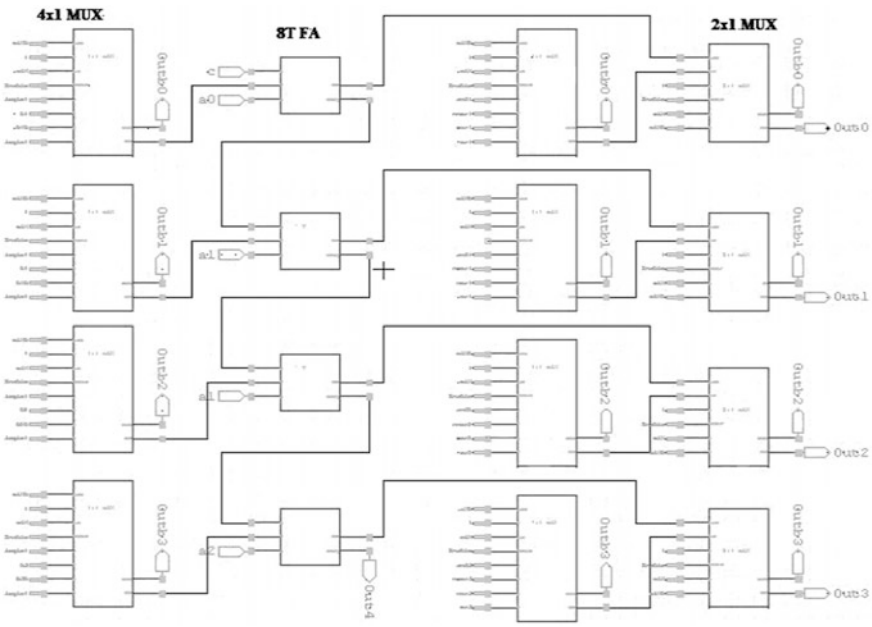


Fig. 8 Implementation of ALU using low-power full adders and data selectors

3.2 Design of 4:1 Data Selector Using Clock Gating

The design of 4:1 multiplexer is shown in Fig. 7. The design of clock gating using multiplexer consumes less power compared to existing designs.

Table 1 Synchronous counters using clock gating

S. No	Synchronous counter without clock gating	Synchronous counter with clock gating
Average power	10.21096 mW	2.375122 mW

Table 2 Design of 4:1 data selector using clock gating

S. No	4:1 data selector without clock gating	4:1 data selector with clock gating
Average power	9.23 μW	5.097 μW

Table 3 Design for ALU using enabled flip-flop gating

S. No	ALU using enabled flip-flop power gating	General ALU
Average dynamic power	0.00221 mW	0.001074 mW

4 Design of ALU Using Enabled Flip-Flop Gating Technique

The design of ALU uses enabled flip-flop gating technique is exposed in Fig. 8. The general design of 4-bit ALU consists of 4:1 data selectors, eight transistor adder circuit, and 2:1 multiplexers.

The average power consumption of synchronous counter, multiplexer, and arithmetic logic unit using enabled flip-flop gating technique was reported in Tables 1, 2 and 3. The simulation results were carried out in 130 μm technologies with a maximum operating frequency of 50 MHz. The dynamic power dissipation of clocked and non-clocked gating techniques are reported in Tables 1, 2 and 3. The latch-based clock gating technique consumes low power consumption compared to conventional approaches.

5 Conclusions

The 3-bit synchronous counter and ALU using clock gating in tanner software were reported. The average power consumption of ALU using multiplexer without clock gating is 3.28 mW and by using multiplexer with clock gating consumes 40% less than the counter without clock gating circuit. The average power consumption of synchronous counter without clock gating circuit is 10.210 mW and the circuit

design using clock gating circuit is 2.375 mW. The design of ALU using multiplexer without clock gating consumes 2.21 μW and the circuit using clock gating consumes 1.07 μW.

Finally, 8T full adder and multiplexer using a synchronous counter with the clock gating technique are used to design ALU and obtained average power of 40% less than existing design.

References

1. Fallah, F., Abdullah, A., Pedram, M.: A robust power-gating structure and power mode transition strategy for MTCMOS design. IEEE Trans. Large Scale Integr. Syst. **15**(1), 80–89 (2007)
2. Howard, D., Shi, K.A.: Sleep transistor design and implementation Simple concept yet challenges to be optimum. In: Proceedings of VLSI-DAT. pp. 1–4 (2006)
3. Jacobson, H., Bose, P., Hu, Z., Buyuktosunoglu, A., Zyuban, V., Eickemeyer, R., Eisen, L., Griswell, J., Logan, D., Sinharoy, B., Tendler, J.: Stretching the limits of clock-gating efficiency in server lass processors. Int. Sympos. High Perform. Comput. Architect. 238–242 (2005)
4. Juan, L.C., Chen, Y.T., Chang, S.C.: An efficient wake-up schedule during power mode transition considering spurious glitches phenomenon. In: IEEE Conference on Computer Aided Design. pp. 777–782(2007)
5. Uyemura, J.P.: Introduction to VLSI Circuits and Systems. Wiley, New York (2002)
6. ChrisjinGnanasuji, G., Maragatharaj, S.: Performance analysis of power gating designs in low power VLSI circuits. IEEE Trans. Comput. **26**, 689–693 (2011)
7. Lee, C., Chang, H., Sapatnekar, S.S.: Full chip analysis of leakage power under process variations, including spatial correlations. Design Automation Conference, pp. 523–528 (2005)
8. Labaey, J.M., Pedram, M.: Low Power Design Methodologies. Kluwer, Norwell, MA (1996)
9. Long, C., He, L.: Distributed sleep transistor network for power reduction. IEEE Transaction on VLSI Systems. **12**(9), 937–946(2004)
10. Long, C., Xiong, J., He, L.: On optimal physical synthesis of sleep transistors. In: Proceedings of ISPD. pp. 156–161 (2004)
11. Nogawa, M., Ohtomo, Y.: A data-transition look-ahead DFF circuit for statistical reduction in power consumption. IEEE J. Solid State Circ. **33**, 702–706 (1998)
12. Pan, Z.D., Ramalingam, A., Devgan, A.: Wake-up scheduling in MTCMOS circuits using successive relaxation to minimize ground bounce. J. Low Power Electron. **3**(1), 25–38 (2007)
13. Rajashekar, P., Malathi, M.: Power efficient multiplexer using DLDFF synchronous counter. IEEE J. Solid State Circ. **42**, 284–289 (2011)
14. Nehru, K., Shanmugam, A., Darmila Thenmozhi, G.: Design of Low Power ALU using 8T FA and PTL based MUX circuits. In: IEEE Conference on Advances in Engineering, Science and Management, 45–49 (2012)

N-bit Pipelined CSM Based Square Root Circuit for Binary Numbers

Siba Kumar Panda, Arpita Jena and Dhruba Charan Panda

Abstract Uniqueness is the key to go forward and maintain a spirited advantage in high-end research. By means of interdisciplinary research from the theory of very large-scale integration to signal processing prospective, it provides a systematic approach to the design and analysis of various circuits that are mainly intended for a variety of VLSI signal processing applications. This work presents an efficient design and implementation of n-bit pipelined square root circuit. An idea of a square root circuit with using a controlled subtract-multiplex (CSM) block is introduced here. In this paper, we have implemented a popular algorithm called non-restoring algorithm for the square root operation of circuits. Anticipated to meet the various design of higher order circuits to use in a range of mathematical operations, this paper provides full exposure to the researchers in the field of VLSI design. Overall, it explores the key themes of designing the square root circuit, its simulation and debugging in Xilinx ISE 14.1 as well as its FPGA implementation

Keywords Squar root circuit · VLSI signal processing · Non-restoring algorithm · CSM · FPGA · VHDL

1 Introduction

The square root [1] of a positive binary number can be determined by using different algorithms. In various signal processing applications, multiplication, squaring, and square root are the majority as well as frequent arithmetic operations.

S. K. Panda (✉) · A. Jena
Department of ECE, Centurion University of Technology and Management, Jatni 752050, Odisha, India
e-mail: panda.sibakumar08vssut@gmail.com

D. C. Panda
PG Department of Electronic Science, Berhampur University, Berhampur 760007, Odisha, India
e-mail: dcpanda@gmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_51

A competent arithmetic manipulation plays an essential task in acquiring the proposed appearance in most of the VLSI signal processing applications. Square root operation has extensive applications in diverse domains of basic sciences and engineering. In this framework, many square root circuits are proposed in the literature. Senthilpari and Kavitha [2], Li and Chu [3], Samavi et al [4], Li et al. [5], O'Leary and Leeser [6] and there also many algorithm proposed for square root circuit design like Newtons method, rough estimation method, digit-by-digit method, and Babylonian method. Senthilpari and Kavitha proposed a design of square root circuit using adders. Samavi et al. offered a classical non-restoring array structure to simplify the circuit without any loss in square root precision. Similarly, O'Leary and Leeser, their design include an adder—substractor and very simple combinational logics.

This has annoyed us to explore the design of n-bit-pipelined CSM-based square root circuit. A method for finding the square root of binary numbers using non-restoring algorithm is presented here. In this work, we have used the non-restoring algorithm to find out the binary square roots [7, 8] for 2-bits, 4-bits, 8-bits, 32-bits as well as for n-bits [9, 10]. This paper describes a way of square root determination of n-bits with simulation results. Basically, NR is a division algorithm which gives two integers "N", "D" and computes the quotients [11, 12]. It is an eminent and prominent method which can be applied in many of the digital circuit designs. The interesting thing in this is that here the every circuit is designed by means of pipelining practice. The NR algorithm is extremely simple to be taught. This algorithm can also be used for the various design of divider architectures [13].

Here, the main motive is to design and implement a pipelined square root circuit for n-bit binary number with a controlled subtract-multiplex (CSM) block. In this work, the n-bit circuit uses the CSM block as the basic building block and for simplification different names are given to the blocks. The number of stages used for n number of input will be its half that is n/2. The number of blocks in 1st level is (n/2) + 2. Each stage is shifted by 2 bits left according to the non-restoring algorithm and the number of CSM block reduces by 1 in each stage. The designed circuit is carried out for VHDL implementation [14]. The main objective is to design, simulate and to implement it in FPGA.

The paper organization is as follows: Sect. 2 describes the non-restoring algorithm and square root of binary numbers. The design of different bits of square root circuit is shown in Sect. 3. Implementation and verification of the design of the n-bit circuit are described in Sect. 4. Section 5 deals with Results and discussions. At end, conclusions are found in Sect. 6.

2 The Non-restoring Algorithm—Square Root of Binary Numbers

A positive binary number's square root can be determined by various algorithms such as Babylonian method, Newton–Raphson method [8], rough estimation method, digit-by-digit method, restoring algorithm [9], etc. But these algorithms are having many limitations like more computational time, more number of operating steps, complexities, etc. These are also not so efficient to be implemented in hardwires and are difficult to provide the exact result. Thus another new algorithm is introduced which is known as the non-restoring algorithm. This algorithm presented in Fig. 1 reduces the arithmetic operations and computational time by skipping the restoring steps. A systematic execution on FPGA is also provided. Thus, a better approach is obtained for square root calculation of positive binary numbers by using the non-restoring algorithm.

The operation of the above flow chart, the non-restoring square root algorithm can be explained by an example:

Let us consider D is an 8-bit positive binary number as the radicand and Q is the squared root of D and R be the reminder. Q will be 4 bit and R of 5 bits. n is how

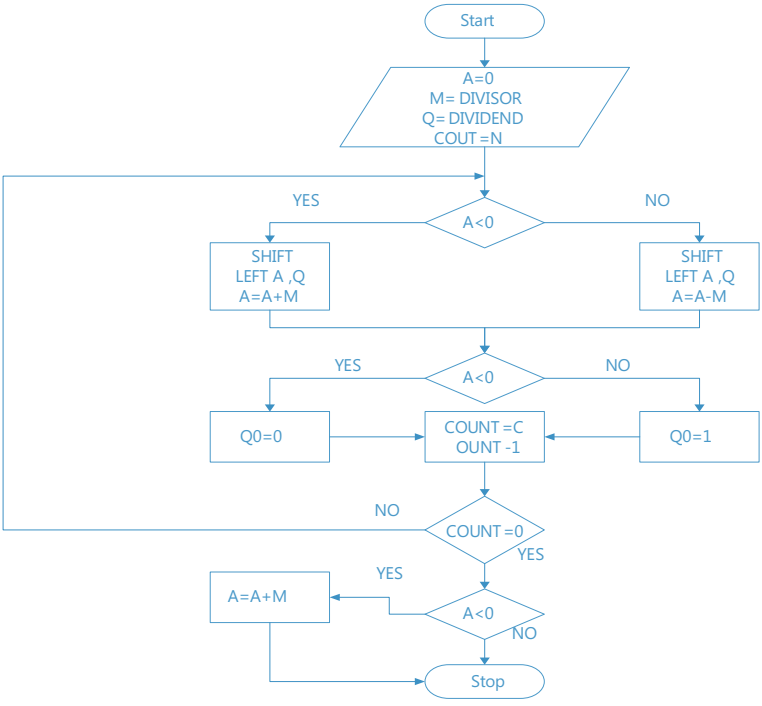


Fig. 1 Algorithmic flow of the non-restoring operation

many times the operation is performed which is same as half of the number of bits of the radicand. So, in this case, it will be from 3 to 0, i.e., 4 iterations.

$D = D_7D_6\ D_5\ D_4\ D_3\ D_2\ D_1\ D_0 = 011111111$ (127)

$Q = Q_3Q_2Q_1Q_0$

$R = R_4R_3R_2R_1R_0$

Set $q_4 = 0$ and $r_4 = 0$ initially.

For n = 3, Q_3 will be 1.

For n = 2, Q_2 will be 0.

For n = 1, Q_1 will be 1.

For n = 0, Q_0 will be 1.

The obtained result is Q = 1011 (11) and R = 00110 (6) for radicand D = 01111111 (127).

3 The Square Root Circuits

An idea used for the square root determination of positive binary number [8, 15] is expressed by means of the circuit diagram. The CSM (controlled subtract-multiplex) block is the basic building structure for the square root circuit. This block uses only subtraction operation. No other arithmetic operations like addition, division, multiplications are required. By using the CSM block the circuitry can be reduced to some extent which results in simple operation. It is shown in the Fig. 2. X, Y, and B are the input signal for the CSM block and b_0 and d_0 are the output signals whereas U is the controlling signal used as an input to the mux unit [16, 17].

A non-restoring pipelined square root circuit diagram of 4-bit is presented [1] in Fig. 3. The number of stages required for the 4-bit circuit is its half. That means two stages are needed. All the blocks used are the CSM blocks. But for simplification, they are named differently. The binary input can be provided by P (P_3, P_2, P_1, P_0) and the output can be obtained at q (q_1, q_0). There is a left shift of 2 bits in every stage. As only the subtraction operation is used, 01 is appended in each stage. If there would be addition operation we have to append 11. If the input is of 4 bit the output result will be of 2 bits (half a bit of input).

We can generalize the non-restoring square root circuit by providing n number of inputs. Any bit value can be applied to the input. The above Fig. 4 represents the n-bit non-restoring pipelined square root circuit. The n-bit circuit uses the CSM block as the basic building block and foe simplification different names are given to the blocks. The number of stages used for n number of input will be its half that is n/2. The number of blocks in 1st level is (n/2) + 2. Each stage is shifted by 2 bits left according to the non-restoring algorithm and the number of CSM block reduces by 1 in each stage. The result is obtained as the output of the D block as U and R is the reminder.

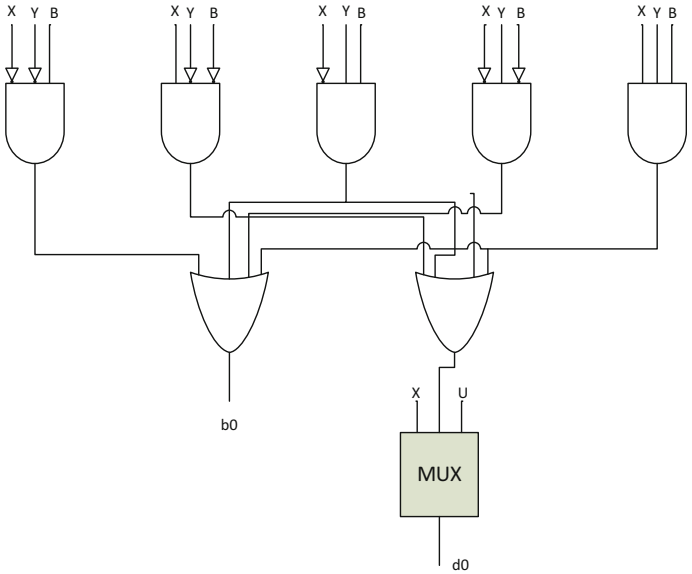


Fig. 2 Basic building block—CSM

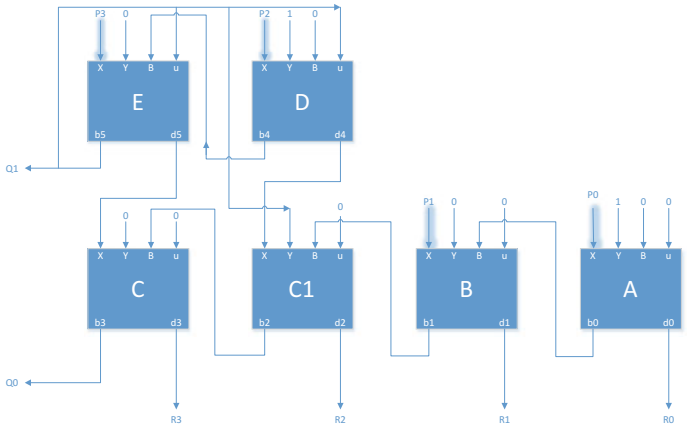


Fig. 3 Block diagram of pipelined 4-bit square root circuit

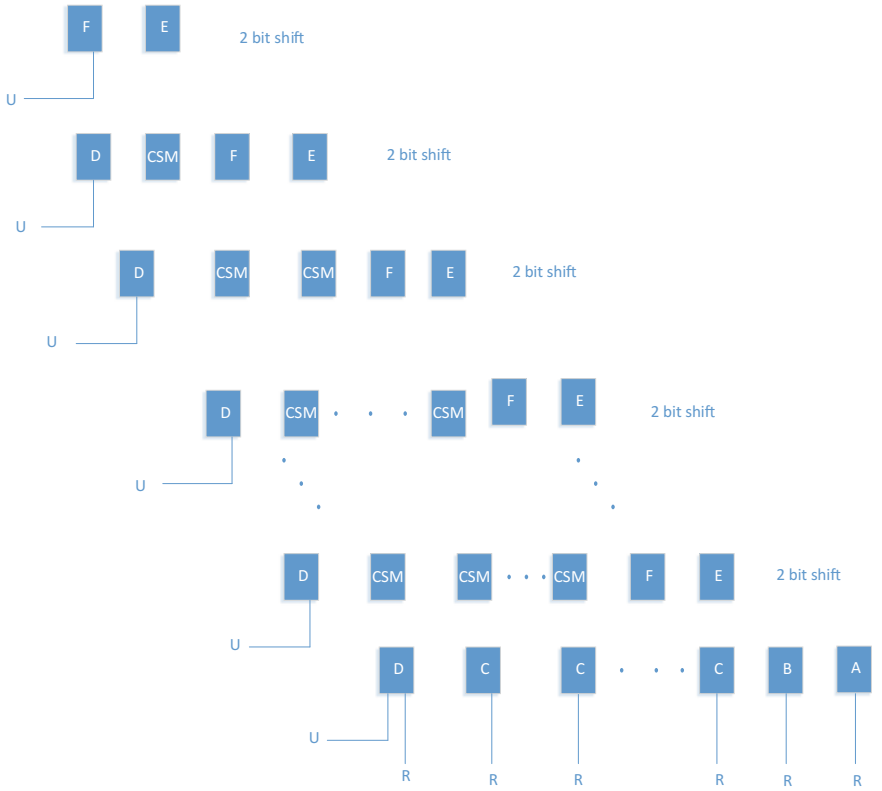
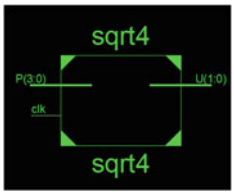


Fig. 4 Block diagram of n-bit pipelined square root circuit

Fig. 5 RTL schematic of 4-bit square root circuit



4 Implementation and Verification

In this exertion, the square root circuit of binary numbers for 4-bit, 8-bit, 32-bit and n-bit [7, 15, 17] is coded in very high-speed IC description language (VHDL) [11]. Then synthesized and pretend in Xilinx ISE 14.1 tool. After that, the designed architecture is put into effect on SPARTAN 3E family board and unscrambled on Spartan 3 XC3S100E. The results are confirmed and found correct.

5 Result Analysis

In this section, the entire simulation result and the test bench waveform output of the designed circuit are given. The Fig. 5 shows the register transfer level (RTL) representation of the 4-bit non-restoring square root circuit with input P (4 bit) and output U (2 bit).

The internal block diagram of the 4-bit square root circuit is shown in Fig. 6, Fig. 7 internal structure and the Fig. 8 represents the test bench waveform output of the circuit. Two random inputs are taken and simulated results are shown. 1001 (9) and 0100 (4) are provided to the input signal p and the simulated squared root output is obtained at U as 11 (3) and 01 (2) respectively. The test bench waveform is obtained as the output which satisfies the square root operation through the designed circuit.

FPGA implementation of the designed circuit is done by using Spartan 3E kit and the hardware output is shown in Fig. 9.

Figure 10 shows the RTL schematic diagram of the 6-bit non-restoring square root circuit which represents the 6-bit input and 3-bit result. The internal block structure of the 6-bit circuit is shown in Fig. 11.

An input for P is taken as 001001 (9) and the required simulation result 011 (3) is the square root of input. The simulation end result of the 6-bit circuit is represented in Fig. 12. The circuit is having 8-bit input as well as 4-bit output. The internal structure is shown in Fig. 13.

The test bench waveform of 8-bit non-restoring square root circuit represents the simulation result in Fig. 14. The 8-bit binary input is provided to the input signal P and the result is obtained in 4 bits. 01000000 (64) is the input and 1000 (8) is the required result.

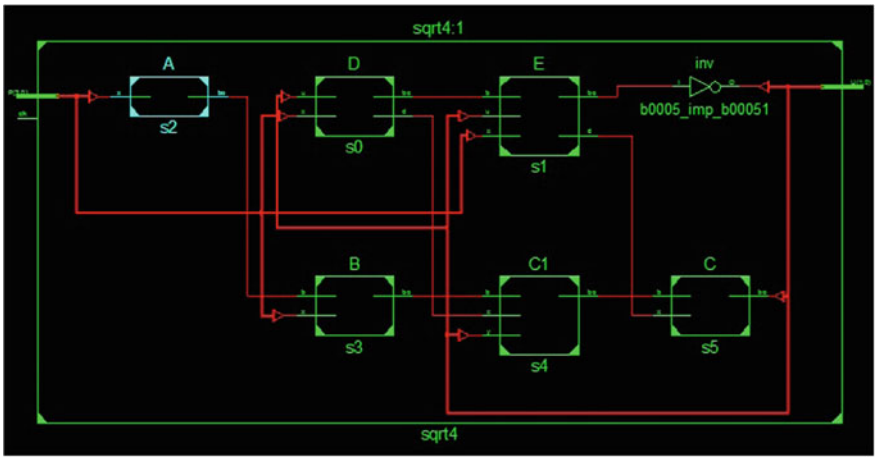


Fig. 6 Internal structure of 4-bit square root circuit

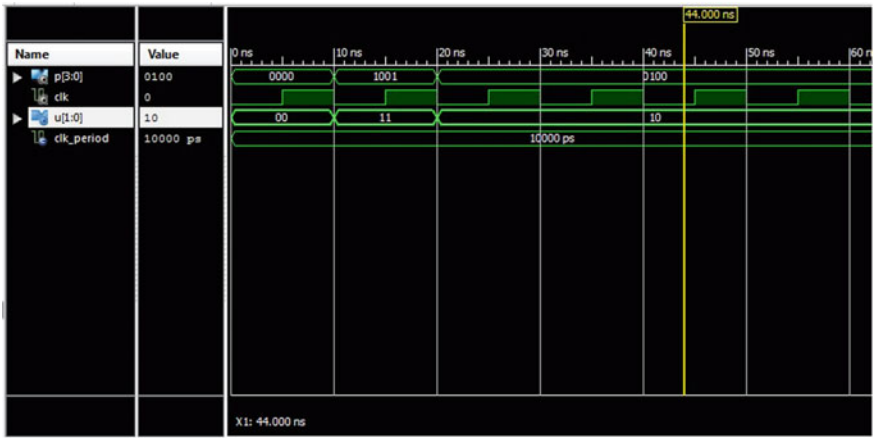


Fig. 7 Simulation end result of 4-bit square root circuit

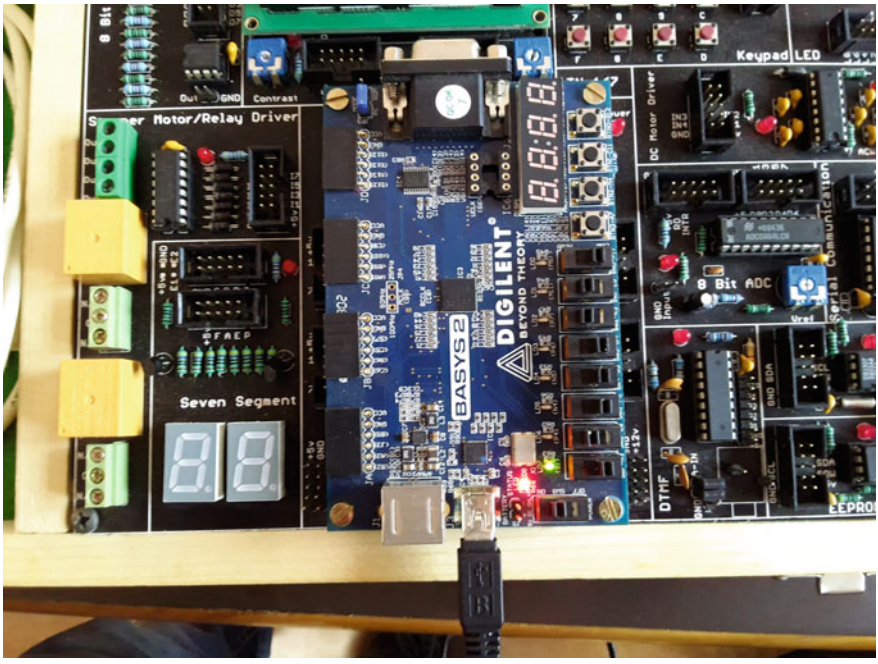


Fig. 8 FPGA implementation of square root circuit

The 16 bit RTL schematic of the designed square root circuit is represented in Fig. 15 with 16-bit input at P and 8-bit output at U.

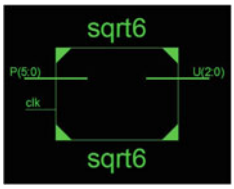


Fig. 9 RTL Schematic of 6-bit square root circuit

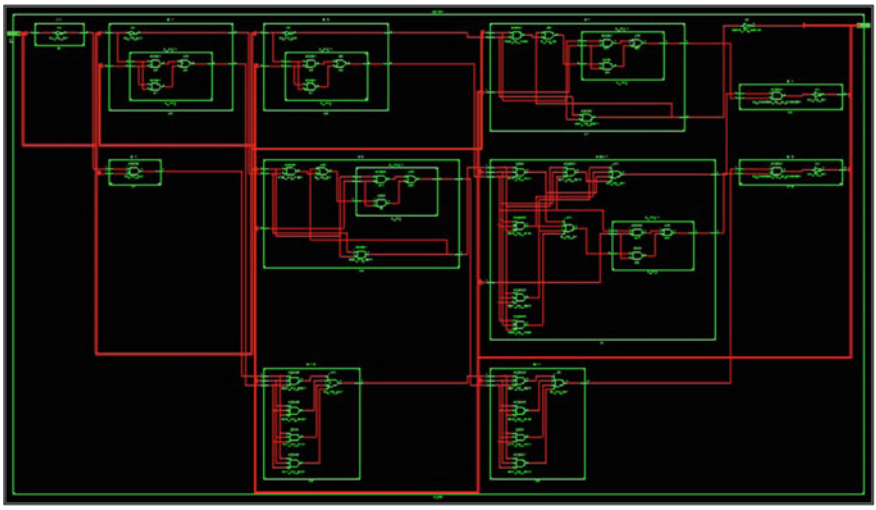


Fig. 10 Internal structure of 6-bit square root circuit

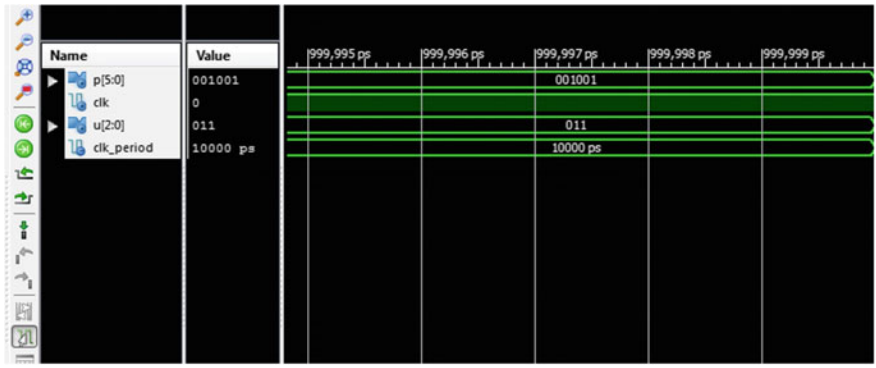


Fig. 11 Simulation outcome of 6-bit square root circuit

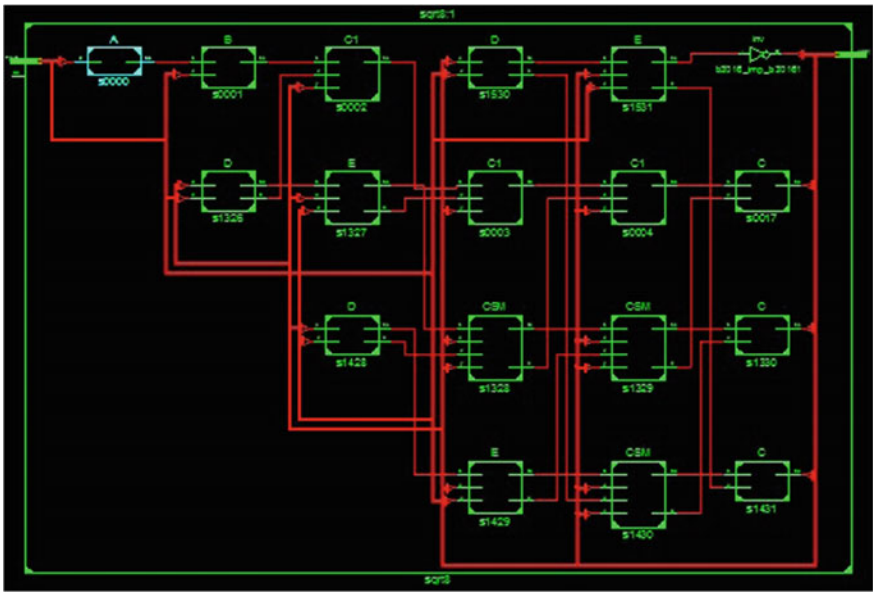


Fig. 12 Internal structure of 8-bit square root circuit

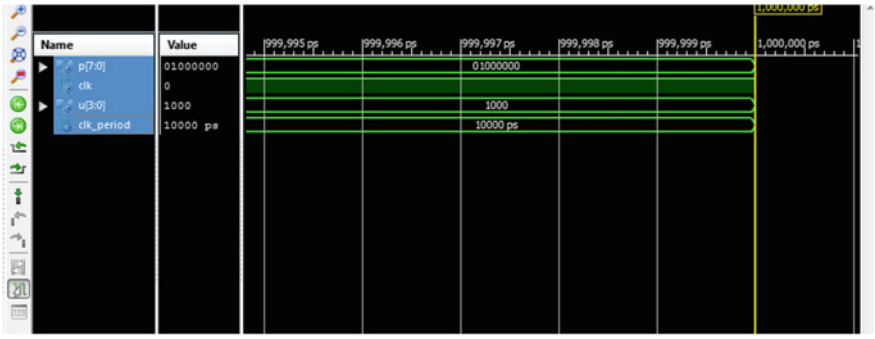
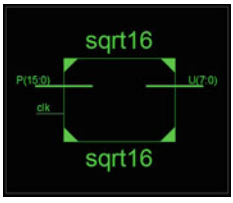


Fig. 13 Simulation result of 8-bit square root circuit

Fig. 14 RTL schematic of 16-bit square root circuit



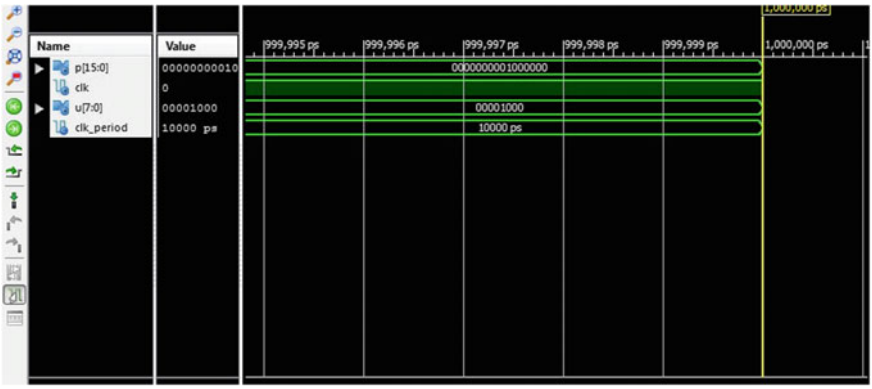


Fig. 15 Simulation outcome of 16-bit square root circuit

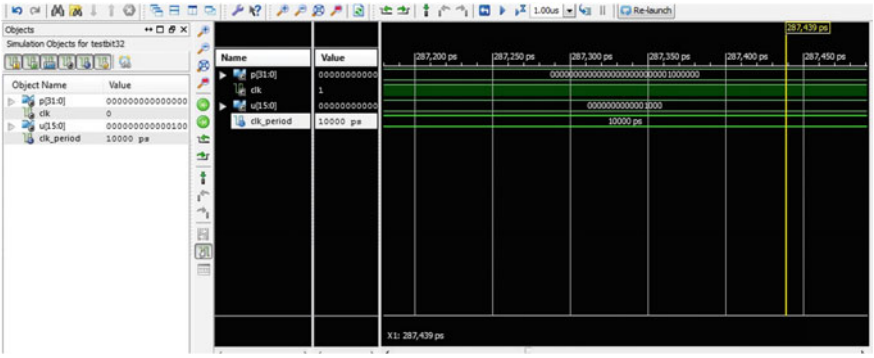
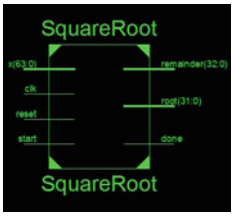


Fig. 16 Simulation result of 32-bit square root circuit

Fig. 17 RTL schematic of n-bit square root circuit



The binary number 0000000001000000 (64) is given to the input P and the square root result is 00001000 (8). This simulation result is shown in Fig. 16 as the test bench waveform of the circuit.

The Fig. 17 shows the test bench waveform output of the non-restoring 32-bit square root circuit which provides a 16-bit simulation output by giving a 32-bit

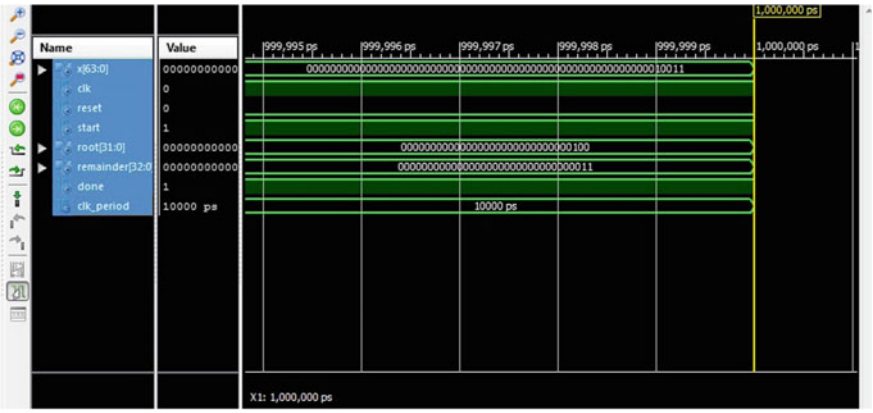


Fig. 18 Simulation outcome of n-bit square root circuit (64-bit input)

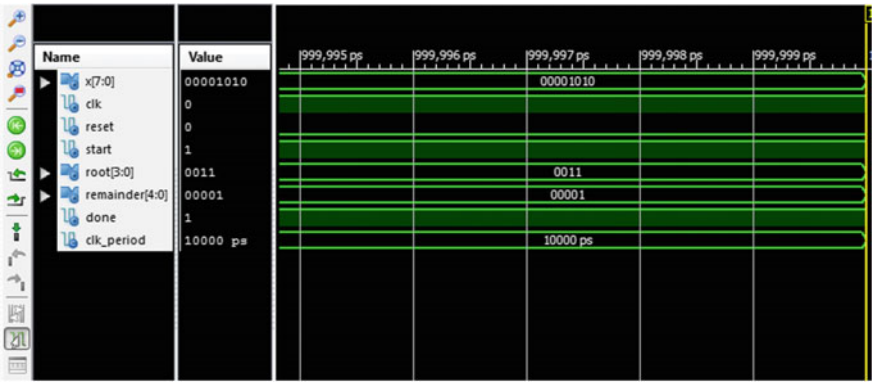


Fig. 19 Simulation outcome of n-bit square root circuit (8-bit input)

input signal. Here 00000000000000000000000001000000 (64) is given as the input and the result is 00001000 (9) obtained at test bench waveform.

Generalized n-bit square root circuit is shown in Fig. 18 with 64-bit input and 32-bit result. The remainder is represented by R. The outcome of the n-bit circuit with 64-bit input is as shown in Fig. 19.

The 64-bit input provided is 0000000000000000000000000000000000000000000000000000000000010011 (19) which results in a root of 32 bit as 00000000000000000000000000000100 (4) and the reminder is 33 bit and is 000000000000000000000000000000011 (3).

Another n-bit non-restoring square root circuit result with 8-bit input and 4-bit output and the remainder is of 5 bit is as shown in Fig. 19. A binary positive number 00001010 (10) is given as input and the root obtained is 0011 (3) of 4 bit and reminder 00001 (1) of 5 bit. We test bench waveform an provide any bit of

binary input to the non-restoring n-bit square root circuit and the result can be obtained as root and reminder in t.

6 Conclusions

The performance of the designed n-bit Square root circuit using non-restoring algorithms proved to be better due to the pipelined structure as well as the introduction of CSM building block. The designed circuit can be implemented in different processors for efficient VLSI signal processing applications. This instigation of research work may assist in the design of higher order digital circuits. The concert of the anticipated method for square root circuit design in terms of simulation results has been presented it shows that it is finer and gives an accurate result. Hence this method may be used for efficient VLSI signal processing applications.

Acknowledgements My sincere thanks to Centurion University of Technology & Management, Jatni, Bhubaneswar, Odisha for providing a high-end research platform.

References

1. Sultana, S., Radecka, K.: Reversible implementation of square-root circuit. In: 18th IEEE international conference on electronic circuits and systems, pp. 141–144. IEEE Press, Canada (2011)
2. Senthilpari, B.C., Kavitha, S.: Proposed low power, high speed, adder-based, 65 nm square root circuit. J. Microelectron. **42**, 445–451, Elsevier Science, (2011)
3. Li, Y., Chu, W.: A new non-restoring square root algorithm and its VLSI implementation. In: IEEE International Conference on computer design, pp. 539–544, Texas, USA (1996)
4. Samavi, S., Sadrabadi, A., Fanian, A.: Modular array structure of non-restoring square root circuit. J. Syst. Architect. 54, 957–966. Elsevier Science (2008)
5. Li, Y., Chu, W.: Implementation of single precision floating point square root on FPGAs. In: 5th IEEE symposium on FPGA for custom computing machines, pp. 226–232. California, USA (1997)
6. O'Leary, J., Leeser, M.: Non-restoring Integer Square Root-A Case Study in Design by Principled Optimization. Technical Report, Cornell University (1994)
7. Sethi, K., Panda, R.: Multiplier less high-speed squaring circuit for binary numbers. Int. J. Electron. **102**, 433–443. Taylor Francis (2014)
8. Rahman, A., Al-Kafi, A.: New efficient hardware design methodology for modified non-restoring square root algorithm. In: International Conference on Informatics Electronics and Vision, pp. 1–6. Dhaka (2014)
9. Sutikno, T.: An efficient implementation of non-restoring square root algorithm in gate level. Int. J. Comput. Theory Eng. **3**, 46–51 (2011)
10. Sajid, I., Ahmed, M.: Pipelined implementation of fixed point square root in FPGA using modified non-restoring algorithm. In: 2nd International Conference on Computer and Automation Engineering, pp. 226–230. IEEE press, Islambad (2011)

11. Wang, L., Schulte, M.: Decimal floating point square root using Newton-Raphson iteration. In: 16th International Conference on Application Specific Systems Architecture Processors, pp. 309–315. IEEE press, USA (2005)
12. Li, Y., Chu, W.: Parallel array implementations of non-restoring square root algorithm. In: International Conference on Computer Design, pp. 690–695. IEEE Press, USA (1997)
13. Panda, S.K., Sahu, A.: A novel vedic divider architecture with reduced delay for VLSI applications. Int. J. Comput. Appl. **120**, 31–36 (2015)
14. Pedroni, V.A.: Circuit design with VHDL. The MIT Press, Cambridge, MA (2008)
15. Sutikno, T., Zakwan, A.: A simple strategy to solve complicated square root problem in DTC for FPGA implementation. In: IEEE Symposium on Industrial Electronics and Application, pp. 691–695. IEEE Press, Penang (2010)
16. Guenther, H.: Arithmetic Operations of the Machine Fundamentals of Digital Machine Computing, Springer publication (1996)
17. Panda, S.K., Jena, A.: FPGA-VHDL implementation of pipelined square root circuit for VLSI signal processing applications. Int. J. Comput. Appl. **142**, 20–24 (2016)

Modelling of a Fibonacci Sequence 8-bit Current Steering DAC to Improve the Second Order Nonlinearities

Anshuman Das Mohapatra and Manmath Narayan Sahoo

Abstract The demand of high speed and high performance IP of current steering DACs are increasing day by day with the advent of telecommunication technologies. The implementation time window is very crucial to launch a new semiconductor product in time. For that reason, prior to transistor level design it is wise to validate the performance of the DAC architecture through system level modelling using MATLAB. This paper elaborates the second order nonlinearities and the capacitive effect of switches through simulations using MATLAB. The Dynamic Element Matching (DEM) helps to overcome static mismatch errors at the cost of degraded Signal-to-Noise Ratio (SNR). The proposed new technique uses a pseudo-randomized DEM method. Rather it is a combination of random selection followed by Fibonacci sequence selection of the unit current sources. This method reduces the implementation time and complexity and improves the SFDR by 4.69 dB at 100 MHz sampling frequency.

Keywords Nonlinearities · Capacitive effect · DEM · Fibonacci DAC MATLAB simulations · SFDR

1 Introduction

Most of the real time signals are analog in nature, but processing of digital data is more advantageous due to their flexibility and re-programmability. Due to the evolution in the digital VLSI technology more and more signal processing is done in

A. Das Mohapatra (✉)
Department of Computer Science and Engineering, National Institute of Technology Rourkela, Odisha 769008, India
e-mail: anshuman.dmp@gmail.com

M. N. Sahoo
e-mail: sahoom@nitrkl.ac.in

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_52

digital domain for the wired and wireless telecommunication systems. As a direct result of that the demand of a high performance DAC has increased over the years. Among the different types of DAC architecture, the current steering DACs are most promising and widely used as they can be easily implemented in CMOS technology and are inherently fast [1, 2]. As the DAC is the first analog block in the transmitter signal path, the overall accuracy of the system depends solely on the performance of the DAC [3]. The current steering DAC is realised by some weighted or unity current sources. The matching of those current sources plays an important role in deciding the overall linearity of the DAC. Like every other device, DACs also suffer from several error sources which modulate the input signal to introduce error in the process of conversion. Dynamic Element Matching (DEM) is a technique that is used to reduce the static mismatch errors of the unity current sources of the DAC [4]. In DEM technique the selection of the current sources is randomized so that the error accumulation is un-correlated with the input signal and the average error is reduced [5]. However, switching the current sources randomly for every sample increases the switching noise and switching powers, which further degrades the spectral performance of the DAC. Hence, DEM method introduces additional power loss, increased complexity and reduces SNR for the DACs. In this paper, the nonlinearities in current steering architecture are modelled in MATLAB with an example of 8-bit DAC. DEM is used to overcome the mismatch errors and to improve the static and dynamic performances of the DAC. Finally, a new design is proposed for enhancing the efficiency of the 8-bit current steering DAC by using Fibonacci sequence. In the proposed method, the CMOS current sources are turned on in a Fibonacci sequence, once the first current source is selected randomly. The proposed architecture has been verified by system level design with various tests carried through MATLAB programming and are supported with the simulation results.

The paper is organised as follows: In Sect. 2 the conventional DAC architecture is discussed followed by highlight on the nonlinearities incorporated by the device through the usage of Simulink in MATLAB. Section 3 presents the idea of DEM to reduce the static errors and the nonlinearities. In Sect. 4 a new architecture is proposed to improve the performance of the DAC with reduced implementation time. The obtained simulation results are recorded and compared with the previous work in Sect. 5. Finally, the conclusion of the work is presented in Sect. 6.

2 General DAC Architectures

Firstly, the general current steering DAC architectures are used to model the 8-bit DAC. The current steering architectures are classified into three major subdivisions, namely, unary, binary and segmented DACs [1].

2.1 Unary Architecture

In this architecture, an equal number of unity current sources (equal to the conversion value of the analog signal) are connected in parallel to produce the continuous time dependent analog output current which is equivalent to the input digital signal [2]. Unary architecture provides most linear output, although the digital area and energy remain an issue for large resolution. The spectral performance of the unary DAC is also superior due to the reduced capacitive switching effects, however, the routing complexity increases for large number of cells cancelling the mentioned advantages. Due to its functioning analogy, this architecture is also known as thermometer coded architecture. For the 8-bit unary DAC modelled using MATLAB, $2^8 - 1$ numbers of unity current sources are used. The current sources are incorporated in a 1-D array. For the digital code 0, all the array elements store the value 0. An input sinusoidal signal is considered as the test input vector to evaluate the performance of the DAC by several specification metrics. A DC shift of A is given to the sine wave to move it entirely in the first quadrant. Each current source is modelled with its equivalent RC model as presented in Eq. (1).

$$T(s) = 1/(1 + sRC) \tag{1}$$

where, $T(s)$ = transfer function, s = variable of Laplace domain, R = resistance of the modelled switches (0.5 Ω) and C = capacitance of the modelled switches (1F). Ilaplace gives an exponential function which is used to model the glitches: $20e^{-20t}$, where t is a time domain variable. As Fig. 1 depicts, for unary architecture, glitches occur mainly in one direction for step jumps in the digital signal.

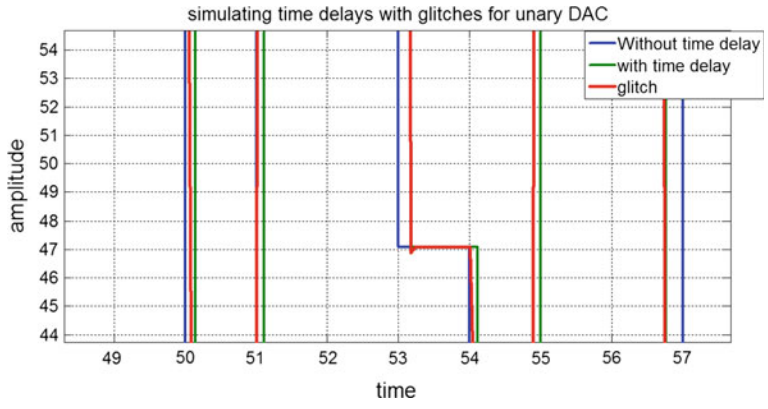


Fig. 1 Time delays and glitches for unary DAC

2.2 Binary Architecture

In this architecture, there is a current source for each converted analog value of the digital signal. Therefore, although the power efficiency for this type of modelling is considerable, but the capacitive switching effect resulting in the glitches is a major drawback for this architecture. As the name denotes, the current sources have magnitudes in the set of all non-negative powers of 2 that is, 1, 2, 4, 8, ... For implementing binary architecture in MATLAB through an 8-bit current steering DAC, an array of 8 current variables is taken with values 1A, 2A, 4A, ..., 128A. They are directly turned on/off by the incoming digital signal value. Glitches occur twice, that is, in both the directions for each step jump in the digital value, which supports the theoretical prediction.

2.3 Segmented Architecture

Segmented Architecture is a combination that takes the advantages of both Binary Architecture and Unary Architecture. A simple representation of this architecture is given in Fig. 3. For most significant bits (MSBs) of the digitized value, unary architecture is used, whereas for least significant bits (LSBs) binary architecture is used. The segmentation ratio depends on the desired accuracy and area trade-offs which has been modelled and discussed in several open literatures [7, 8]. Out of 8-bits, the 4 LSB bits are processed using the previously defined binary architecture ($I_0 = 1$A, I_1,$= 2$A, I_2,$= 4$A and I_3,$= 8$A). The first 4 MSB bits are processed using the previously defined unary architecture (240 unity current sources). For these two tasks, two separate loops were used in MATLAB. For a segmented DAC, glitches occur twice for each step jump for the binary handled bits and once for the unary handled bits. Although these general current steering DAC architectures for an 8-bit DAC solve the purpose, but the INL/DNL errors modulate the original signal beyond the acceptable limits. The SFDR obtained in each of the test cases suggests poor functionality of the device. In order to study the effect of harmonics for 8-bit segmented DAC, Simulink library is used. For decoding the incoming digital bits into binary and unary a structure as shown in Fig. 2 is used.

2.4 Modelling Segmented DAC

For modelling the nonlinearities of the 8-bit current steering DAC using SIMULI-NK, an analog source, quantizer, scopes, zero-order hold, FFT block and spectrum analyser is required as shown in Fig. 4. The successive incremental and

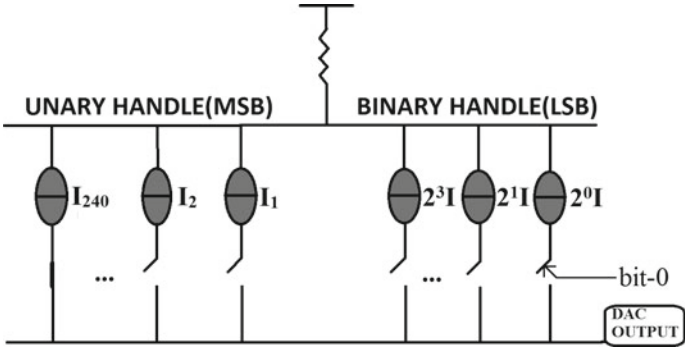
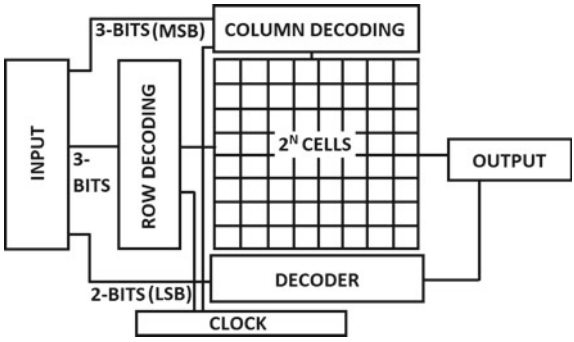


Fig. 2 Segmented DAC architecture

Fig. 3 Decoding layout



decremental approach gives rise to stairs, an intermediate form between digital signal and analog signal, which is given in Fig. 5. The Zero-Order Hold block is used to capture the required FFT plot with desired number of sample inputs, whose output is observed and studied through the spectrum analyser. Fig. 6 presents the output of spectrum analyser. The symmetric behaviour about the central frequency (CF) depicts the effect of the presence of higher order harmonics in the output signal. For improving the 8-bit DAC and to make it more reliable during mismatch environments, Dynamic Element Matching technique is incorporated in the 8-bit DAC designed.

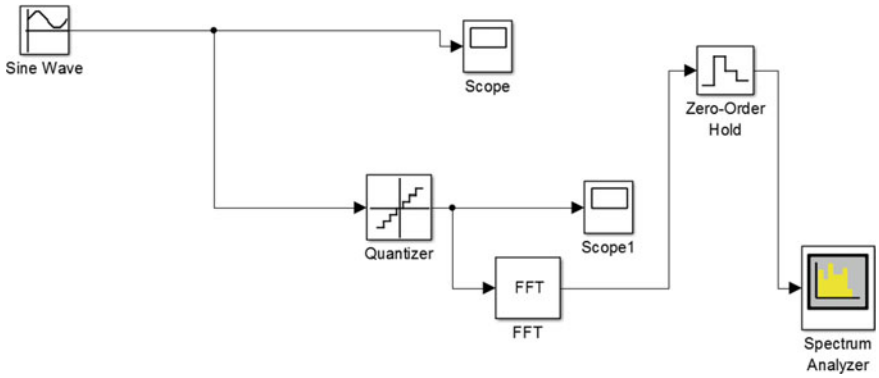


Fig. 4 Block diagram of segmented DAC architecture

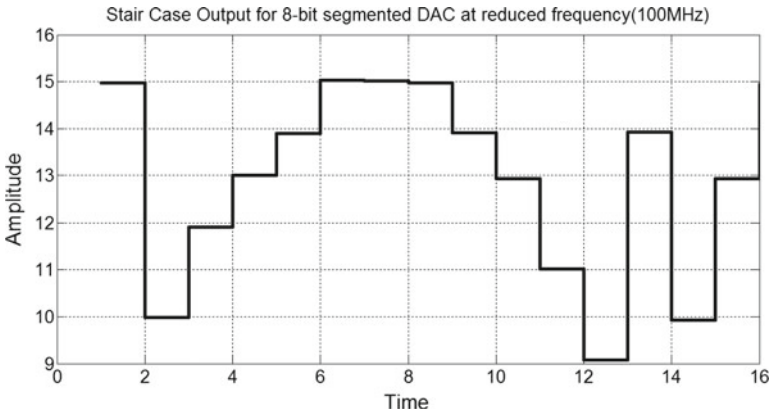


Fig. 5 Stair case DAC output

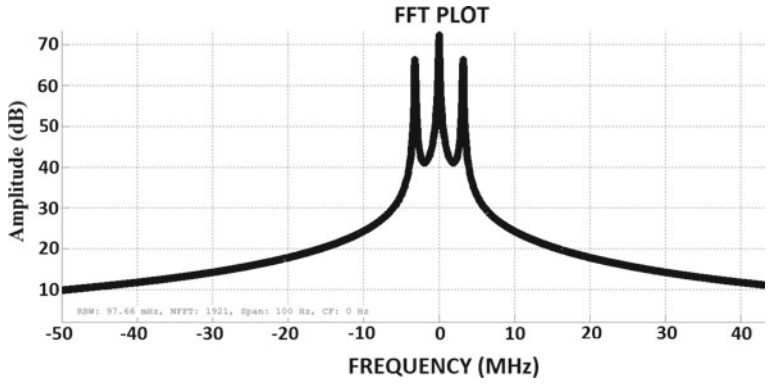


Fig. 6 Harmonics arising from DAC processing

3 Dynamic Element Matching To Model Fault Tolerant DACs

In segmented or unary architecture, it is quite practical that one or more unity current sources may become unreliable for the given task. Therefore, the choice of selection of the unity current sources should be randomized which provides an improved averaged output. For the Fast Fourier Transform (FFT) plot in the frequency domain, the earlier equation used is modified in the following way for reducing the nonlinearities and achieving more practical plots.

$$f_{\text{in}} = M_{\text{c}} * f_{\text{s}}/N \tag{2}$$

where, f_{s} = sampling frequency (100 MHz), N = bits of information, f_{in} = signal frequency and M_{c} = an odd positive integer, most preferably a prime number (127), within the limits of the maximum bits of information processed by the DAC [8].

This modified algorithm avoids the successive adding of the corresponding harmonics, and also shifts the FFT plot away from the origin towards the positive axis. The FFT plot for DEM DAC is given in Fig. 7. The prime number (127) is chosen as the value of M_c on an experimental basis through various tests. The room for improvement lies in the fact that not all the selections need to be randomized. The proposed Fibonacci architecture explores this concept for achieving better test results from the modelled DAC architecture.

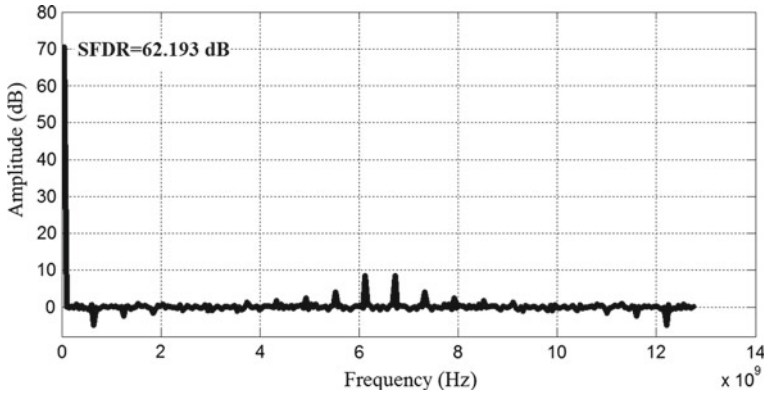


Fig. 7 FFT plot for DEM DAC

4 Proposed Fibonacci DAC Architecture

In this section, the proposed 8-bit Fibonacci DAC architecture is discussed. The proposition uses DEM technique as the parent concept. The Fibonacci sequence is itself a varying interval pattern, that is, the jump between the entities is not fixed. For segmented DACs, instead of triggering the current sources in a complete randomized way, the choice of the first current source under unary architecture can be chosen randomly from a CMOS circuit. The subsequent unity current sources must be chosen from a circular array where the numbers are assigned starting from the first chosen current source numbered as 0. Thereafter, only the current sources that form a Fibonacci pattern are selected for the incoming digital bits. Since, for the 2nd, 3rd, 4th, …, current sources, we are choosing from an established fixed pattern, the running time of this algorithm is better. Based on the number of comparisons, the proposed algorithm incurs much lesser running time overhead than for the DEM architecture.

Pattern establishment requires defining a circular array with Fibonacci numbers, the last number being just greater than 2^8, for the 8-bit Fibonacci DAC: {0, 1, 1, 2, …, 377}. Let D denote the 8-bit digitized value to be converted to analog form. Let UD denote the bits to be processed under the unary architecture and BD denote the same under the binary architecture. UD_d is the decimal value of the MSB bits of D. The BD bits directly control the binary current sources represented by IB. However, the UD bits do not affect the selection of unity current sources denoted by IU. The first element of IU, let IU_i is triggered by a random function. The random value once obtained, R_{in}, say, acts as a pseudo-random number generator. Every time it is incremented by one and a current source corresponding to next Fibonacci Sequence number identified by the *FibCurEle* data structure, is selected. The process continues till the required number of unity current sources under unary architecture are chosen.

Algorithm 1: Proposed Fibonacci DAC Algorithm

Data: D=($b_7b_6b_5b_4b_3b_2b_1b_0$): The digitized code.
UD=($b_7b_6b_5b_4$) BD=($b_3b_2b_1b_0$) UD_d=$(value)_{10}$
// The 8-bit digitized value D is to be converted to analog form.
Result: A for each digital bit.
Initialization: *FibCurEle*[]={0,1,1, …,377};*IB*[]={$2^0, 2^1, 2^2, 2^3$};*IU*[]={1,1, …,1 (upto $2^8 - 1$)};*NOAuc*=0;

```
for each BD_i ∈ BD do
    IB[i]=1;
end
for each UD_i ∈ UD do
    R_in =Random number between 0 to 377 both inclusive
    IU[FibCurEle[R_in]%IU.length]=1;
    NOAuc=NOAuc+1;
    repeat
        R_in = R_in + 1;
        R_in = R_in%FibCurEle.length
        IU[FibCurEle[R_in]%IU.length]=1;
    until (++NOAuc<UD_d);
end
```

5 Simulation and Result

In order to establish the feasibility of the proposed 8-bit Fibonacci Sequence DAC, MATLAB programming environment is used. The implemented unary and binary DAC are integrated together with the proposed algorithm. A MATLAB main script is prepared which calls the user- defined functions and presents the results in the form of plots. A user-defined function "switchonrf.m" is programmed according to Algorithm 1.

The DNL/INL errors (Figs. 9 and 10) for each digitized input code are obtained by using the following formulae.

$$DNL(i) = (I(i+1) - I(i))/(L_s) - 1 \tag{3}$$

where, L_s = Ideal LSB step.

$$INL(i) = \sum_{i=0}^{N} DNL(i) \tag{4}$$

The results are compared for the usual 8-bit current steering DEM segmented DAC and the proposed 8-bit current steering Fibonacci segmented DAC along with the previous work in this area and are given in Table 1. The FFT plot of Fibonacci DAC (Fig. 8) gives a high value of SFDR which suggests good performance.

It can be easily observed that the proposed Fibonacci Sequence DAC outperforms the Dynamic Element Matching DAC and the Fibonacci DAC [9]. The DNL error is comparable to that of DEM DAC since we are still using randomization in the form of pseudo-random numbers, but in an efficient way. The probability of selecting a completely different current source under unary architecture is almost same in both the cases. However, the fixed Fibonacci pattern saves considerable computational

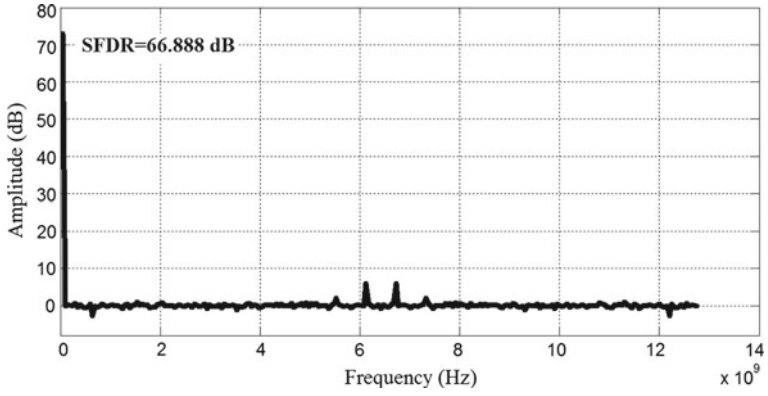


Fig. 8 FFT plot for Fibonacci DAC

Table 1 Performance comparison

	Fibonacci DAC	DEM DAC	REF [9]
DNL Error [LSB]	±0.2056	±0.2031	±0.25
INL Error [LSB]	±0.0298	±0.0270	±0.22
SFDR [dB]	66.888	62.193	50

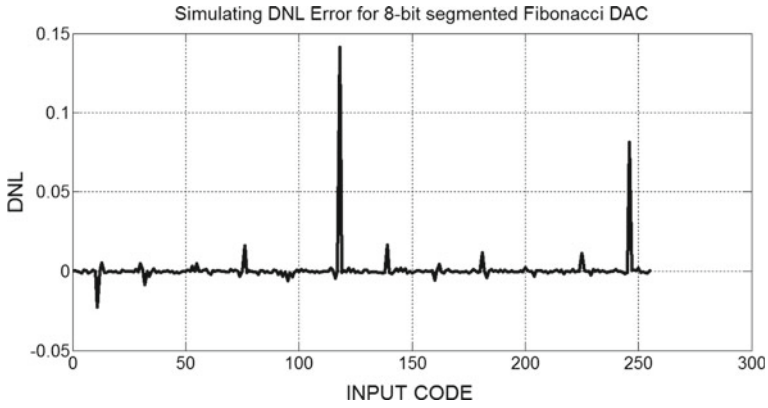


Fig. 9 DNL error for fibonacci DAC

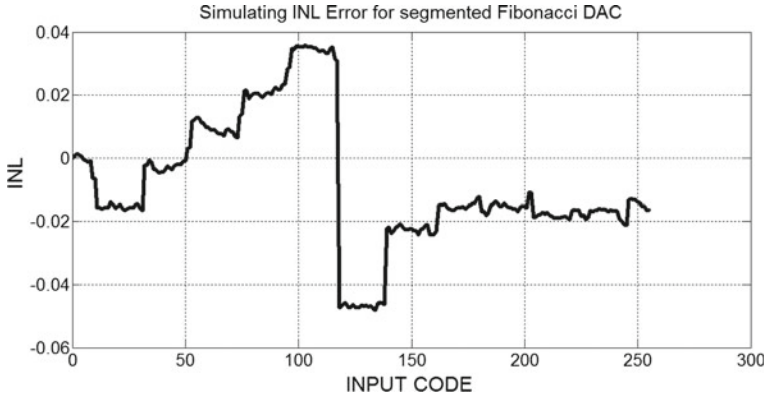


Fig. 10 INL error for fibonacci DAC

time. The proposed algorithm is therefore, fast and uses minimum runtime overhead. It does not use a complex decoding logic like that of [9] but enhances the benefits of DEM through the Fibonacci pattern. Hence, it shows improved SFDR and acceptable INL error values.

6 Conclusion

In this paper we have proposed a Fibonacci sequence based current steering DAC. The system level modelling of the DAC is done in MATLAB. The proposed system level model of the DAC in MATLAB will help the designer to reach a quick decision to select the suitable architecture depending on the targeted specifications. The complexity of the proposed DAC architecture is O(n) in the average case, and takes much lesser operational time compared to DAC architecture with DEM algorithm. The DNL and INL errors of the proposed Fibonacci DAC are almost comparable with the conventional 8-bit current steering DAC with DEM algorithm and much better than the previous work in this field [9]. Although, the 8-bit Fibonacci DAC shows improved SFDR in simulation. High SFDR is required to realise the DACs in the telecommunication transceivers. The proposed 8-bit Fibonacci current steering DAC with the improved spectral performances can be used in transceivers for efficient and accurate digital communication. The glitches due to the capacitive nature of the switches could be minimized further by incorporating more bits in unary architecture.

References

1. Bosch, A.V.D., Steyaert, M., Sansen, W.: Analog Integrated Circuits and Signal Processing, An Accurate Statistical Yield Model for CMOS Current-Steering D/A Converters, December 2001, **29**(3), pp. 173–180
2. Lin, C., Bult, K.: A 10-b, 500-MSample/s CMOS DAC in 0.6 mm^2, IEEE J. Solid-State Circ. **33**, 12, December (1998)
3. Sarkar, S., Banerjee, S.: An 8-bit 1.8V 500MSPS CMOS Segmented Current Steering DAC . In: IEEE Computer Society Annual Symposium on VLSI (ISVLSI), 268273, May (2009)
4. Gerasta, O.J.L., Catane, J.R.M., Mortel, M.J.O. : 8-bit DAC with Partial Randomization Dynamic Element Matching for Nonlinear Distortion Correction. 1-6, 2012 IEEE Region 10 Conference
5. Shen, W.: Transistor Level Implementation of Data Weighted Averaging (DWA). May (2010)
6. Razavi, B.: RF Microelectronics. Prentice Hall, Inc., ISBN 0138875715 (1998)
7. Zhu, B., Song, Z., Yang, D., Ye, Y., Li, F.: A 8-bit 200MSmaple/s CMOS DAC. pp. 198 –200, IEEE International Conference on Anti-Counterfeiting, Security and Identification (2011)
8. Bosch, A.V.D., Borremans, M.A.F., Steyarert, M.S.J., Sansen, W.: A 10-b 1-GSample/s Nquist Current-Steering CMOS D/A Converter. IEEE J. Solid-State Circ. **36**, 315–324, Mar-(2001)
9. Hokazono, K., Kanemoto, D., Pokharel, R., Tomar, A., Kanaya, H., K. Yoshida, A Low-Glitch and Small-Logic-Area Fibonacci Series DAC, pp. 1-4, IEEE 54th International Midwest Symposium on Circuits and Systems (MWSCAS) (2011)

Design of Low-Power and High-Performance Network Interface for 2 × 2 SDM-Based NoC and Implementation on Spartan 6 FPGA

Y. Amar Babu, G. M. V. Prasad and John Bedford Solomon

Abstract As VLSI technology is growing exponentially, silicon chips can accommodate more cores on a chip and this will lead to very high computational power but poor communication among on-chip processors and memory. To overcome this, we proposed spatial division multiplexing based network-on-chip with the modified network interface. Proposed network interface provides high throughput with the optimized area and consume very low power. We have evaluated proposed SDM-based NoC (network-on-chip) with high-performance network interface for 2 × 2 network which occupied only 4% of resources on Xilinx Spartan6 SP605 FPGA. We modeled the network interface using VHDL and multicore platform is prepared by using Xilinx EDK and verified computationally complex application at 88.6 MHz processor frequency but achieved high throughput.

Keywords Network-on-Chip (NoC) · FPGA (Field programmable gate array) Spatial division multiplexing (SDM)

1 Introduction

The state of art in chip multiprocessors design in nanometer technology regime is to accommodate more processors, memory cores, and complex functional controllers onto a single silicon chip [1]. This huge integration of IP cores onto a single chip

Y. Amar Babu (✉)
Department of ECE, LBR College of Engineering, Mylavaram, India
e-mail: amarbabuy77@gmail.com

G. M. V. Prasad
Department of ECE, B.V.C Institute of Technology & Science, Batlapalem, AP, India
e-mail: drgmvprasad@gmail.com

J. B. Solomon
Department of ECE, Karunya University, Coimbatore, Tamil Nadu, India
e-mail: beford.solomon@gmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_53

gives challenging problems in on-chip communication architectures to chip designer. Copper interconnects are not fit for today's on-chip communication infrastructure because of high current density, electron migration, and signal integrity issues. One alternative solution for on-chip communication infrastructure problems is network-on-chip. In network-on-chip designer can choose a variety of topologies like mesh, tree, hybrid network and it can be circuit switching type or packet switching type [2–5]. Varieties of network-on-chip architectures have been proposed in recent years. Mainly NoC architectures are classified based on the link between routers those can be Time Division Multiplexing (TDM) or Spatial Division Multiplexing. In TDM-based NoC at every router, buffers are required for routing packets to adjacent routers and for flow control mechanism to maintain error-free transmission between adjacent routers [6–8]. To optimize buffers requirement at every router of NoC SDM based NoC can be opted which will send data in a serialized fashion to the adjacent router and multiple data can be sent to destination.

The architecture of network-on-chip as shown in Fig. 1. It consists of router and network interface as basic building blocks. The router is connected to adjacent IP core through network interface and router is also connected to the adjacent router through spatial division multiplexing links. Based on switching type and multiplexing type we have different router architectures and network interface microarchitectures. Router mainly consists of crossbar switch and virtual channels and buffers which are used to hold and forward packets to adjacent routers and local IP core. Every router has five ports east, west, north, south and local. IP core is connected to NoC through a local port of routers. NoC architectures are based on Globally asynchronous and locally synchronous (GALS) this feature enhances the scalability and flexibility of network [9].

Today's FPGAs are more complex and it gives ultimate solution in speed, low power, reuse when compared with ASIC (Application-Specific Integrated Circuits)

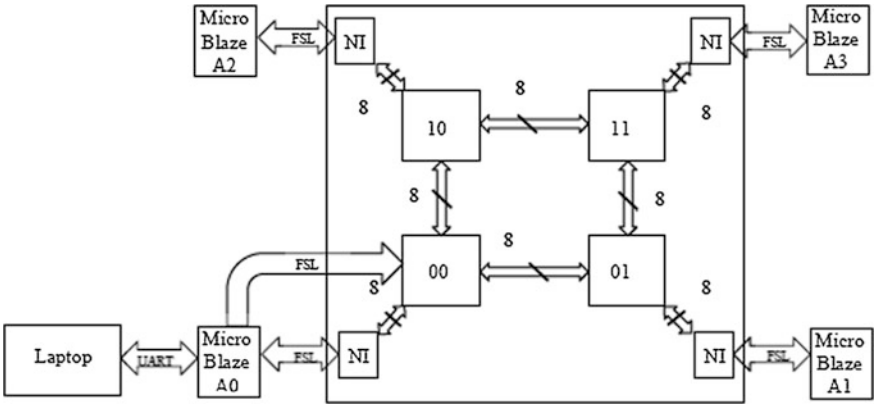


Fig. 1 Network-on-chip

Design. Network-on-chip architectures can be implemented on FPGAs to meet time to market design metric and reconfigurability. In this paper, we have proposed SDM based NoC with high-performance network interface to achieve high performance for a computationally complex application which has to consume low power in real time [10]. We have used Xilinx Spartan6 FPGA to evaluate our proposed NoC and verify various complex application performances.

2 Buffer-Less SDM-Based NoC

To improve performance and optimize area we have modeled NoC router without buffers. Proposed buffer less SDM-based NoC has two layers which are responsible for programming the network to fix router to router links and forwarding the packets from source node to a destination node in a serialized format. Mesh type topology has been selected because it has flexibility and scalability features. The basic architecture of proposed buffer less SDM based NoC is as shown in Fig. 2. Figure 2 shows 2 × 2 NoC best for cryptography and image compression applications. It controls network layer which is used for fix links between the router in order to transfer data from source node to destination node. Every router has five ports and two channels, one is for sending data serially and other one is receiving data serially.

3 NoC Design Methodology

Proposed NoC architecture basic building blocks, NoC router and network interface are modeled using VHDL. NoC router has five ports and each port size has two 8-bit wide in and out pins which are used to send and receive data to and from other adjacent routers. The architecture of NoC router is as shown in figure. Each side of NoC router has 3-bit select lines to select specific input port pin and output port pin to transmit and receive data serially. Based on application fix link between adjacent routers. In our experimental work, we have selected 2 × 2 mesh topology and node 00 is connected to node 01, node 01 is connected to node 11 then node 11 is

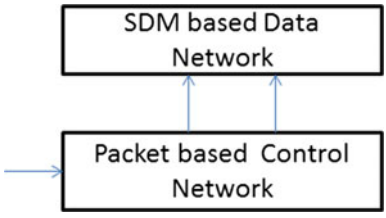


Fig. 2 Physical view of NoC

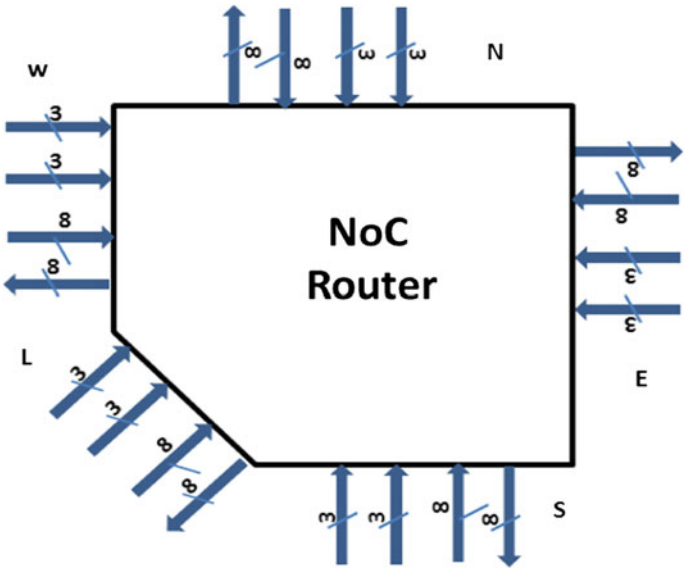


Fig. 3 NoC router

connected to node 10 and finally data sent from node 10 to node 00. For every node one MicroBlaze soft core is connected to local port of the router. Here node means router. Node 00 is selected to send programming data for fixing link between adjacent routers. Once fixing nodes links data layer is responsible for sending data from MicroBlaze soft core to other node soft cores or memory cores (Fig. 3).

Control network and data network have been modeled using VHDL and then top level 2 × 2 NoC architectures are implemented using Xilinx ISE tools. The top-level NoC platform used as IP core in EDK tool to prepare MPSoC (multi-processor system-on-chip) platform and then hardware platform which consists of NoC, 4 MicroBlaze soft cores and shared PLB bus has been exported into Xilinx SDK (Software Development Kit) to develop application software using C/C++ language. In our prototype, we have developed various applications like AES algorithm, JPEG compression, JPEG 2000 Compression and verified on the prepared MPSoC platform.

4 High-Performance Network Interface

Network interface consists of serializer, deserializer, in_FSM, out_FSM, setup_FSM_NI, these blocks are used to get data from IP core in 32-bit format and send to NoC network through a router. So network interface act as bridge between NoC router and IP softcore. It will receive data from IP core in 32-bit format and

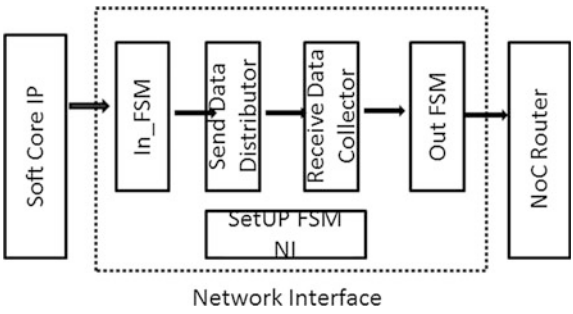


Fig. 4 Network interface

then serializes sent to NoC router and receive serialized data from NoC then deserializes to make 32-bit format for IP softcore. Based on port size of router number of serializers, deserializer, in_FSM, out_FSM, will be decided. In 2 × 2 NoC experimental setup size of router port is 8 bit and a number of channels are two, send data channel and receive data channel. So one serializer, one deserializer, 8 in_FSMs, 8 out_FSMs are required for high-performance network interface. For test setup 4 network interface blocks are required to build 2 × 2 SDM based NoC and 9 FSL (Fast Simplex Link) are required for 2 × 2 NoC. One FSL links are used for sending programming data from PC to node 00 of control network and for every MicroBlaze two FSL links are used to connect to NoC (Fig. 4).

5 Results

The proposed 2 × 2 SDM based NoC with high-performance network interface has offered better performance when compared with existing techniques. It occupied only 4% of resources on Spartan6 SP605 FPGA and consumes low power in the order of mW because of customized hardware in network interface. Proposed techniques use 32-bit to 1-bit serializer instead of 32 bit to m bit converter which saves up to 50% hardware and consumes only 2% power. The synthesis report of the proposed architecture is as shown in Table 1 which provides about LUT (Look Up Table) utilization and Slices occupied and Flip Flops used on Spartan6 FPGA. Table 2 provides information about various complex applications like the performance of AES algorithm, JPEG compression, JPEG 2000 Compression. In experimental setup, application is divided into tasks and distributed among 4 microblaze soft processors. Execution time and power consumption of each application were compared with other competitive architectures like TDM based NOC, hybrid NoC and shared PLB bus. Proposed architecture provides superior execution time and power consumption (Fig. 5).

Table 1 Synthesis report of test setup

Device utilization summary			
Logic utilization	Used	Available	Utilization (%)
Slice registers	6062	54576	11
LUTs	5257	27288	19
LUT-FF pairs	2457	8862	27
IOBs	288	296	97
BUFG/BUFGCTRLs	2	16	12

Table 2 Performance analysis

Application	Proposed NoC architecture	SDM based NoC	TDM based NoC	PLB shared bus	AXI architecture
AES	2s	4.5s	6s	59s	50s
JPEG	3s	6.3s	8s	90s	80s
JPEG2000	4s	7s	9s	120s	90s

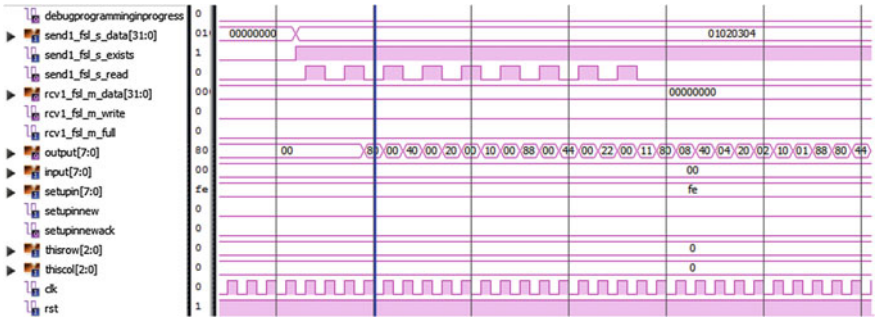


Fig. 5 Simulation of network interface

6 Conclusions

In this paper, a novel high-performance network interface for spatial division multiplexing based network-on-chip has been proposed and evaluated performance by prototyping on Xilinx Spartan6 SP605 FPGA. Our technique occupies only 4% of resources on target FPGA and 50% area optimized when compared with network interface with 32 bit to m bit serializer and consumes only 2% of power when compared with existing techniques. This work can be extended by considering flow control algorithms at the router and by increasing flexibility of NoC router.

References

1. International Technology Roadmap for Semiconductors: Semiconductor Industry Association, Dec 2015
2. Dally, W.J., Towles, B.: Route packets, not wires: on-chip interconnection networks, In: ACM/IEEE Design Automation Conference (DAC), June 2001
3. Benini, L., De Micheli, G.: Networks on chips: a new SoC paradigm. Computer **35**(1), 70–78 (2002)
4. Hemani, A., Jantsch, A., Kumar, S., Postula, A., Oberg, J., Millberg, M., Lindqvist, D.: Network on chip: an architecture for billion transistor era. In: IEEE NorChip Conference, Nov 2000
5. Havemann, R.H., Hutchby, J.A.: High performance interconnects: an integration overview. Proc. IEEE **89**(5) (2001)
6. Amar Babu, Y., Prasad, G.M.V.: Performance analysis and implementation of modified SDM based NoC for MPSoC on Spartan6 FPGA. Int. J. Res. Eng. Technol. (IJRTE) (2016)
7. Amar Babu, Y., Prasad, G.M.V.: Design and implementation of area and power efficient network on chip on FPGA. Int. J. Electron. Eng. Res. (IJEER) (2014)
8. Amar Babu, Y., Prasad, G.M.V.: An area and power efficient on chip communication architectures for image encryption and decryption. Int. J. Res. Eng. Technol. (IJRET) **3**(5) (2014)
9. Amar Babu, Y., Prasad, G.M.V.: Implementing a next-generation design: 3D-IC design. Int. J. Electron. Eng. Res. (IJEER) **3**(3), 343–349 (2011)
10. Bjerregaard, T., Sparso, J.: A router architecture for connection oriented service guarantees in the MANGO clockless network-on-chip. In: Proceedings of the Design, Automation and Test in Europe, DATE'05, pp. 1226–1231 (2005)

Aspects of Machine Learning in Cognitive Radio Networks

Harmandeep Kaur Jhajj, Roopali Garg and Nitin Saluja

Abstract Radio devices employ the technology of cognition to sense the radio frequency environment. Cognitive radios are mainly for reliable communication and for efficient usage of spectrum resources. They identify an apt action that needs to be applied to a particular situation called reasoning. The results of applied actions provide information which helps them to modify their behavior called learning. Learning can be categorized as *supervised learning and unsupervised learning*. In supervised learning, the radio device already has some knowledge about the surrounding environment. In unsupervised learning, the radio device has no knowledge of the surrounding environment. To design the real cognitive systems, various AI techniques are metaheuristic algorithms, rule-based systems, artificial neural networks, ontology-based systems, hidden Markov models and case-based systems. This paper focusses on the learning aspect of the CR and the role of artificial intelligence in CR.

Keywords Cognitive radio (CR) · Cognitive engine (CE) · Artificial intelligence (AI) Secondary user (SU) · Primary user (PU) · Spectrum holes

1 Introduction

Cognitive radio is a technology that is used to efficiently utilize the spectrum band. The primary user of the spectrum uses TDMA (time division multiple access) or FDMA (frequency division multiple access) techniques to access the spectrum.

H. K. Jhajj (✉) · R. Garg
UIET, Panjab University, Chandigarh, India
e-mail: harmandeepk517@gmail.com

R. Garg
e-mail: roopali.garg@pu.ac.in

N. Saluja
Chitkara University, Patiala, Punjab, India
e-mail: nitin.saluja@chitkara.edu.in

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_54

With these techniques, some holes in the spectrum remain unused [1, 2]. So to efficiently use the spectrum, these unused holes can be provided to the secondary user until the holes are idle [3]. For this, some intelligent technology is needed that can sense the unused bands and can take the actions that how these bands can be allocated to the secondary user. Cognitive radio is the best technology for this purpose. Cognitive radio uses the AI (Artificial Intelligent) techniques to perform its tasks [4]. The main tasks of the cognitive radio are to sense its RF environment and gather the information of the environment, to take the actions for the situations and to gather the knowledge for the results of the actions and uses this knowledge to modify the behavior of the CR so that the efficiency can improve [5]. This process of gathering the knowledge is called learning [5]. There are two types of learning—unsupervised learning and supervised learning [5]. The main part of the CR is CE (cognitive engine). It is the heart of CR. The main cognitive tasks of CR, i.e., learning and reasoning are performed by CE [4]. CE can work independently or can be worked with multiple processes or multiple CEs. Game theory is used for the interactions between multiple CEs [4].

In this paper, various AI techniques are discussed that can be used in CR to behave intelligently. Section 2 presents the learning and reasoning aspects in CR, various algorithms of unsupervised and supervised learning. Section 3 describes techniques of AI and gives the conclusion of this paper.

2 Learning in CR

Learning is the process of modification of the behavior of the agent by gathering the knowledge about the environment and by observing the previous actions of the agent [6, 7]. Learning is the main component of the CR. The learning process is based on gathering the knowledge. Also, cognition means the technique of knowing something [6]. To know something, knowledge is a must [8]. Hence, in learning and cognition, knowledge is the common part [5]. The relationship between the intelligence, learning, and cognition is depicted in Fig. 1. By learning, the agent can organize the information, classify the information into the categories, and also generate and generalize the information [7]. With learning, CR also performs reasoning and awareness. All these components of the CR can be shown by a cognition cycle [5] which is shown in Fig. 2.

The main difference between the learning and reasoning is that learning considers the past and present observations about the environment while reasoning only considers the current observations about the environment to take actions [5]. It does not consider the past history.

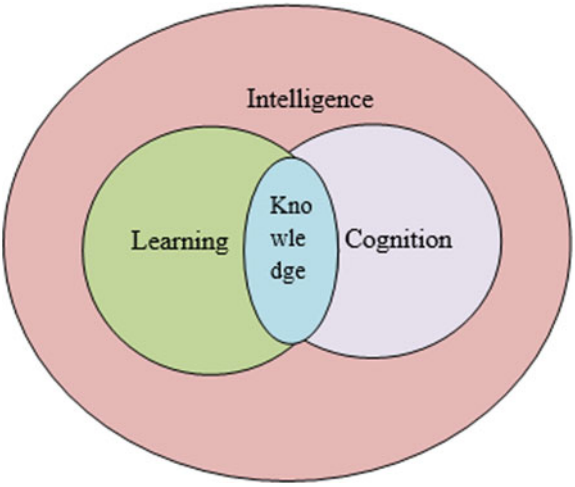


Fig. 1 Relationship between intelligence, learning, and cognition

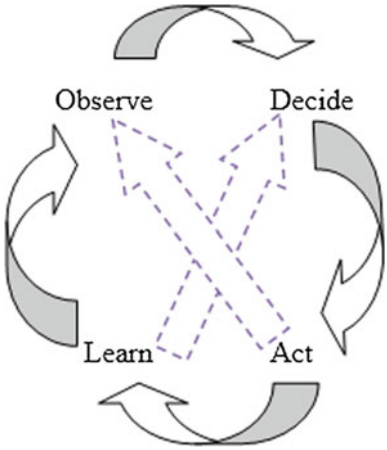


Fig. 2 Cognition cycle

2.1 Unsupervised Learning

Unsupervised learning is the learning having no knowledge about the environment [9]. The various unsupervised learning algorithms are as follows:

Reinforcement learning is the technique that works in an autonomous environment. A feedback is given to the agent after executing each action [10]. RL uses this feedback to modify the behavior of the agent. There are two methods of RL: *Trial and Error and Delayed Reward* [11].

The working of the RL is shown in the Fig. 3. In the given figure, a_t is the action that has to be taken at time t. This action is taken by considering the observation of the state variable s_t and reward r_t. r_t is the feedback of the action taken at time $t - 1$,

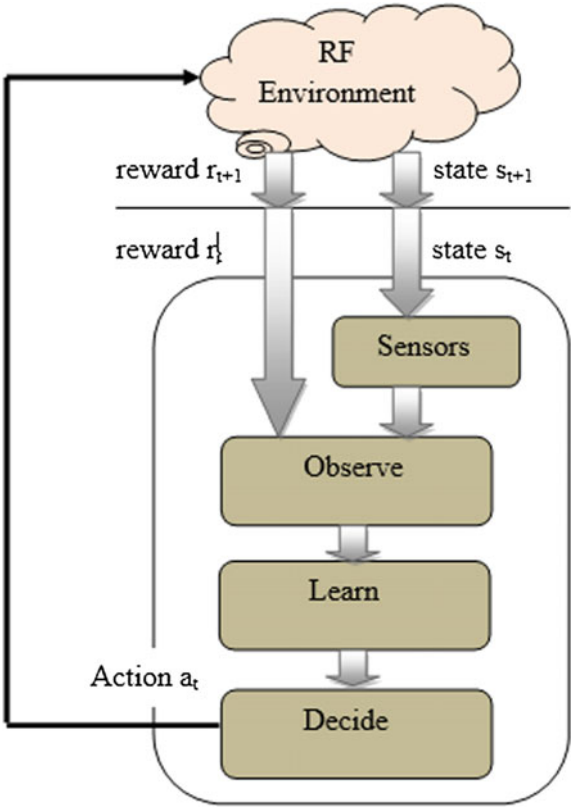


Fig. 3 Working of reinforcement learning [13]

i.e., the action a_{t-1} which changes the state from s_{t-1} to s_t. Now, action a_t will give the reward r_{t+1} and will change the state from s_t to s_{t+1} [12].

Game theory-based learning. This learning is used in the environment where multiple CRs tries to learn the environment to adapt their behavior [5]. In this type of environment, centralized control cannot be used as it will increase the cost and time of communication. Complexity will also increase with centralized control. So, game theory-based learning is the best solution for these types of environments [5].

According to the game theory, there are several entities in the environment that are called players. Each player has many available actions and an utility function [14]. A player decides the value of its utility function by observing the actions of other payers. Every player tries to maximize its utility function [5].

Threshold learning. When a CR is implemented in the mobile devices, several problems occur for the CR to adapt the environment due to the mobility of the mobile devices and switches of the transmission between channels [15]. When moves from place to place, noise or interference be encountered. Due to this noise or interference, it is difficult for the CR to fit in its environment completely and hence the performance is degraded.

In such situations, threshold learning can be used by CR to adapt to its environment. Threshold learning does the continuous learning from the past experience for the dynamic adaptation of the parameters of CR so that the system performance can improve [16]. By observing the effects of the past actions on the system performance, threshold learning optimizes the values of the parameters of CR to obtain the desired performance [5]

2.2 *Supervised Learning*

Supervised learning is the learning when there is some prior knowledge available about the environment. There are several types of supervised learning as discussed below:

ANN. Artificial neural network is the network having various neurons connected to form the network [5]. It has the ability to gather the knowledge from the environment using learning process and to store the information for future use [17]. Interneuron connections are used to store the information.

SVM. Support vector machine is a technique that is used for object classifications and pattern recognition [5]. The strategy of the SVM is that it classifies the objects into various classes according to their features. Objects having the same features are put into the same class and objects having different features are put into the different classes [18]. The vector which describes the features of the object is called feature vector. It can be of any dimension depending upon the number of features in the feature vector. Hyperplanes are formed to classify the classes.

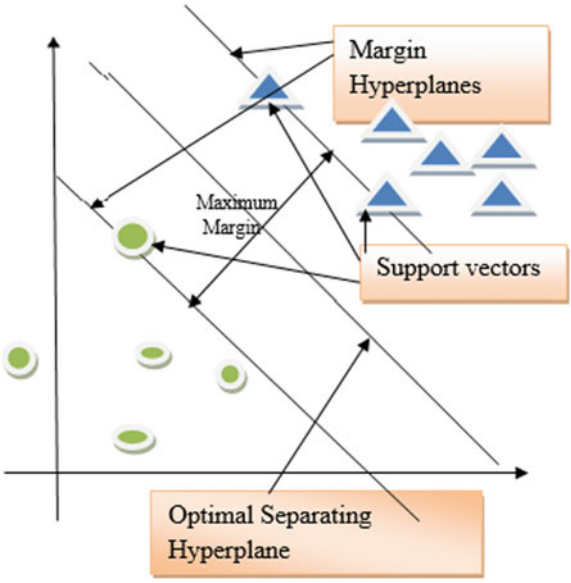


Fig. 4 Basic idea of SVM

Hyperplanes distinguish the different classes from each other. SVM is used to classify the signals. Figure 4 gives the basic idea of SVM.

In Fig. 4, circles and triangles are the two classes of the objects. Hyperplanes that are present at the boundaries of the classes are called margin hyperplanes and the hyperplane that is present at the highest margin is known as optimal separating hyperplane. Objects present on the hyperplanes are called support vectors [5].

3 Conclusion

CR is a technology that is used to efficiently utilize the spectrum band. Learning is the process of modification of the behavior of the agent by gathering the knowledge about the environment and by observing the previous actions of the agent. There are two types of learning: unsupervised learning and supervised learning. In unsupervised learning, the agent has no knowledge of the environment while in supervised learning, the agent already has some knowledge of the environment. Reinforcement learning, game theory-based learning, Threshold learning are the types of unsupervised learning and ANN, Support vector machine is the types of supervised learning. Artificial Intelligence is the intelligence exhibited by the machines or software. ANN, metaheuristic algorithms, HMM, RBS, OBS, CBS are the techniques of AI that are used by CR.

References

1. Ubiquisys: Ubiquisys and Percello Unveil Next Generation Femtocell Platform. http://www.ubiquisys.com/ub3b/pressreleases.php?id=115
2. Greenis, P.: Smart communications for smart grids In: Nanotech Conference Proceedings Expo (2009)
3. Zhu, X.-L., Liu, Y.-A., Weng, W.-W., Yuan, D.-M.: Channel sensing algorithm based onneural networks for cognitive wireless mesh networks In: 4th International Conference Proceedings WiCOM, pp. 1–4 (2008)
4. He, A., Bae, K.K., Newman, T.R., Gaeddert, J., Kim, K., Menon, R., Morelas-Tirado, L., Neel, J., Zhao, Y., Reed, J.H., Tranter, W.H.: A survey of artificial intelligence for cognitive radios. IEEE Trans. Veh. Technol. **59**(4), 1578–1592 (2010)
5. Bkassiny, M., Li, Y., Jayaweera, S.K.: A survey on machine-learning techniques in cognitive radios. IEEE Commun. Surv Tutor. **15**(3) (2013)
6. Le, B., Rondeau, T.W., Bostian, C.W.: Cognitive radio realities. Wirel. Commun. Mobile Comput **7**(9), 1037–1048 (2007)
7. Mitchell, T.M.: Machine Learning. McGraw-Hill, New York (1997)
8. Mitola, J.: Cognitive radio for flexible mobile multimedia mommunications In: IEEE International Workshop Proceedings on Mobile Multimedia Communications, pp. 3–10 (1999)
9. BBN Technologies: XG Policy Language Framework. http://www.autonomiccommunication/org/web/bodies/DARPA_XG_rfc_policylang.pdf
10. Paine, J.: Expert systems. http://www.j-paine.org/students/lectures/lect3/lect3.html

11. Yau, K.-L. A., Komisarczuk, P., Teal, P. D.: Applications of reinforcement learning to cognitive radio networks In: IEEE International Conference on Communications Workshops (ICC), pp. 1–6 (2010)
12. Venkatraman, P., Hamdaoui, B., Guizani, M.: Opportunistic bandwidth sharing through reinforcement learning. IEEE Trans. Veh. Technol. **59**(6), 3148–3153 (2010)
13. Jiang, T., Grace, D., Mitchell, P.: Efficient exploration in reinforcement learning-based cognitive radio spectrum sharing. IET Commun. **5**(10), 1309–1317 (2011)
14. Unnikrishnan, J., Veeravalli, V.: Cooperative sensing for primary detection in cognitive radio. IEEE J. Sel. Topics Signal Process. **2**(1), 18–27 (2008)
15. Han, Y., Pandharipande, A., Ting, S.: Cooperative decode-and-forward relaying for secondary spectrum access. IEEE Trans. Wirel. Commun. **8**(10), 4945–4950 (2009)
16. Clancy, T., Khawar, A., Newman, T.: Robust signal classification using unsupervised learning. IEEE Trans. Wirel. Commun. **10**(4), 1289–1299 (2011)
17. Han, Z., Zheng, R., Poor, H.: Repeated auctions with bayesian nonparametric learning for spectrum access in cognitive radio networks. IEEE Trans. Wirel. Commun. **10**(3), 890–900 (2011)
18. Ganesan, G., Li, Y.: Cooperative spectrum sensing in cognitive radio. IEEE Trans. Wirel. Commun. **6**(6), 2204–2213 (2007)

FPGA Implementation of Buffer-Less NoC Router for SDM-Based Network-on-Chip

Y. Amar Babu, G. M. V. Prasad and John Bedford Solomon

Abstract Transistor size shrinking day-to-day as technology node moves towards deep sub nano meter node so, Interconnects dominate overall performance of system-on-chip. Conventional shared bus architecture could not handle on-chip communication issues like bandwidth, power consumption, and signal integrity. To overcome these issues network-on-chip provides the best alternative to shared bus architectures. In this paper, novel NoC router has been proposed to minimize area design metric to 50% and power consumption for efficient on-chip hybrid communication in the spatial division multiplexing-based network-on-chip. Using proposed NoC router a 2 × 2 SDM-based NoC has been implemented on Xilinx Spartan 6 FPGA and performance are evaluated and compared with TDM-based NoC architectures.

Keywords Bufferless NoC router · SDM based network-on-chip
TDM based network-on-chip

1 Introduction

Today's chip multiprocessors have more than 100 cores to meet application needs like high bandwidth, low power consumption. The on-chip communication plays a vital role to meet design metrics. The conventional shared bus architecture is obsolete to meet multimedia application needs from system-on-chip. Network-on-chip solves

Y. Amar Babu (✉)
Department of ECE, LBR College of Engineering, Mylavaram, India
e-mail: amarbabuy77@gmail.com

G. M. V. Prasad
Department of ECE, B.V.C Institute of Technology & Science, Batlapalem, A.P, India
e-mail: drgmvprasad@gmail.com

J. B. Solomon
Department of ECE, Karunya University, Coimbatore, Tamil Nadu, India
e-mail: beford.solomon@gmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_55

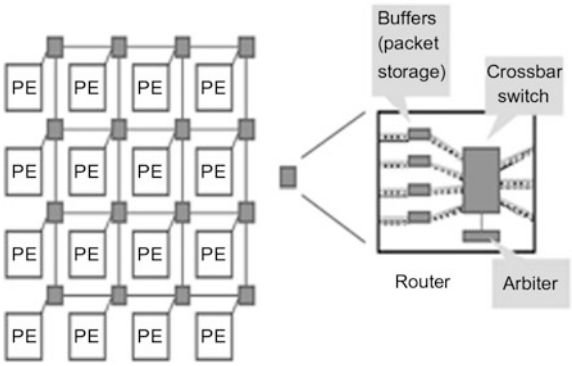


Fig. 1 Network-on-chip

problems of on-chip communication issues. The basic network-on-chip as shown in Fig. 1. NoC provides flexibility and scalability for many core chips and provide globally asynchronous and locally synchronous feature and supports packet switching, circuit switching and hybrid switching to handle data traffic in on-chip communication [1–4]. NoC supports multiple clocking schemes to maintain synchronization among core, memory and control blocks. In recent research on network-on-chip unveils time division multiplexing-based NoC, spatial division multiplexing-based NoC and hybrid NoCs for a variety of applications. The basic network-on-chip architecture has routers which have multiple ports to send and receive packets from adjacent routers and local soft cores. In these architectures network interface is another basic building block which acts as an interface between soft cores and NoC routers, decides the performance of top-level architecture. The design network-on-chip for any application-specific architectures depends on network topology, switching technique, routing algorithms, flow control algorithms. To optimize the performance of the application on-chip multiprocessor designer has to choose best topology like mesh, torus, cube, switching technique like packet switching, circuit switching and hybrid switching, routing techniques like deterministic and nondeterministic, flow control techniques like token flow control, positive acknowledge and negative acknowledgment flow control [5–7]. All these design parameters decide design metrics of network-on-chip architectures. In this paper a novel NoC router based on SDM-based NoC has been proposed, implemented on FPGA, evaluated performance in the next sessions.

2 Buffer-Less Network-on-Chip

The physical structure of network-on-chip as shown in Fig. 2. It consists of the dual layer for communication among soft cores, memory and control blocks. Bottom layer forwards packets to program the network and fixed physical links between NoC routers and top layer forwards actual data packets among softcore, memory

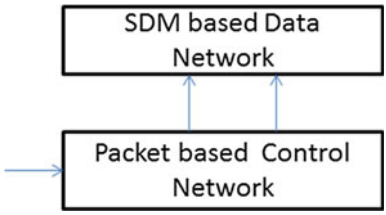


Fig. 2 Abstract view of network-on-chip

blocks, and control blocks. In the first stage of design 2 × 2 NoC architecture which uses mesh topology and can be extended to maximum 7 × 7 size depends on resources required on FPGAs. In this architecture packet switching is used for control network layer to program and fix links between adjacent routers circuit switching is preferred for actual application data transfer between soft cores on-chip multiprocessors. These physical links between routers can be static or dynamically fixed based application demand and memory-on -hip. Dynamic configuration requires extra memory but can provide better performance [8–10]. The entire architecture blocks are modeled using VHDL and programming data for control network can be sent through application program which is coded in C language. The size of physical links between routers can be changed based on application bandwidth requirement. In the test setup, the links between routers are 8 bit in size can be 16 bit or 32 bit but requires extra resources on FPGA. To integrate user-defined IP NoC on to Spartan 6 FPGA, Xilinx EDK tool has been used to prepare hardware platform. The prepared hardware platform using Xilinx EDK has proposed NoC and four MicroBlaze soft processors, nine fast simplex links and one shared processor local bus.

3 Buffer-Less NoC Router

For the proposed 2 × 2 SDM-based NoC requires 4 routers for mesh topology. The top-level structure of NoC router is as shown in Fig. 4. It has five ports east, west, south, north, and local port, each port has one input port of size 8 bit, one output port of size 8 bit, 3 bit allocated input, and 3 bit allocated index. These physical links between routers are spatial division multiplexed and data sent from source node to destination node serially and desterilized at the receiver side of the network interface. So, to send data from any port of five ports of router serially to adjacent router through the use of allocated input and allocated index ports. Using programming data these links will be programmed initially based on application demands. Here proposed network has four routers and each router has a unique number like 00, 01, 10, and 11. To evaluate the performance of AES algorithm, JPEG compression node 00 is connected to node 01, node 01 is connected to node

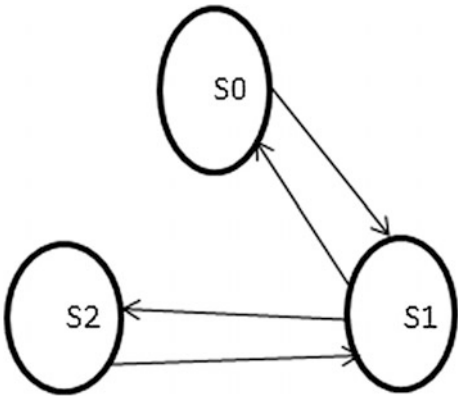


Fig. 3 State diagram of setup FSM for NoC router

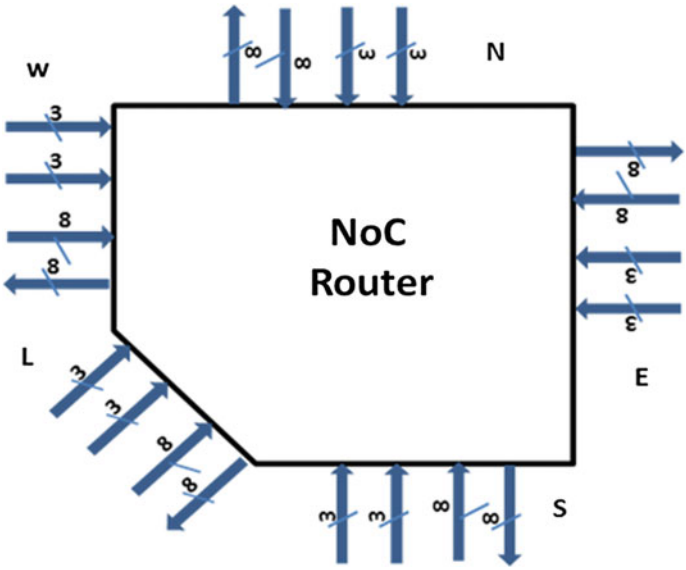


Fig. 4 Buffer less NoC router

11, node 11 is connected to node 10 and node 10 is connected to node 00. Programming data to fix links between nodes is first sent from node 00. To program these links router has setup FSM which as three states. The state diagram of setup FSM is as shown in Fig. 3.

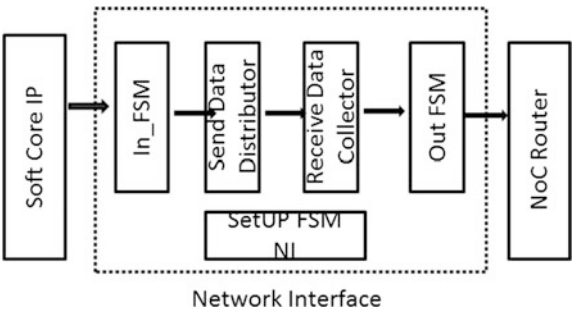


Fig. 5 Network interface

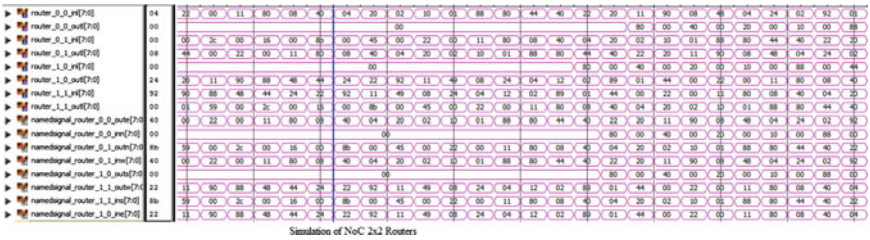


Fig. 6 Simulation of 2 × 2 NoC router

4 Network Interface

The network interface for proposed SDM-based NoC has transmitter side and received side as shown in Fig. 5. In every side, data send and received serially from softcore to NoC router. The network interface consists of send data distributor, receive data collector, input FSM, output FSM and setup FSM. In present architecture 32 bit to 1-bit serializer is proposed instead of 32 bit to m bit converter which minimizes the area required for the network interface on FPGA.

5 Results

The experimental set up has 2 × 2 network-on-chip, 4 network interface, 4 MicroBlaze soft processors. Simulation results of NoC router is as shown in Fig. 6 which shows four routers control data, fast simplex link data, allocated input, allocated input index data. It shows how data is transmitted from source node 00 to node 01, node 10, node 11. The synthesis report of the test setup is as shown in Table 1 which gives how many resources are used and available on FPGA and percentage utilization. Table 2 shows application performance in execution time and power consumption and compared with standard NoC benchmark architectures and shared bus architectures.

Table 1 Device utilization summary

Synthesis report summary on SP605			
Logic utilization on SP605	Used	Available	Utilization (%)
Slice registers	6062	54576	10
Look up tables	5257	27288	18
Look up tables–flip-flops pairs	2457	8862	26
Input output blocks	288	296	95
BUFG/BUFGCTRLs	2	16	11

Table 2 Performance analysis on SP605

Application	Proposed NoC architecture	SDM-based NoC	TDM-based NoC	PLB shared bus	AXI architecture
AES	2s	4.5s	6s	59s	50s
JPEG	3s	6.3s	8s	90s	80s
JPEG2000	4s	7s	9s	120s	90s

6 Conclusions

On-chip communication architectures improve the performance of the overall system-on-chip. In this paper, a novel NoC router has been modeled using VHDL and implemented on FPGA. Performance of proposed architecture is evaluated and compared with other standard benchmark architectures. Area of NoC router has been optimized to 50% and thereby power consumption also get optimized. Overall performance of proposed architecture shows better results and this work can be extended by including flow control algorithms in NoC router level which may require extra FPGA resources.

References

1. Amar Babu, Y., Prasad, G.M.V.: Implementing a next-generation design: 3D-IC design. Int. J. Electron. Eng. Res. (IJEER) **3**(3), 343–349 (2011)
2. Amar Babu, Y., Prasad, G.M.V.: Design and implementation of area and power efficient network on chip on FPGA. Int. J. Electron. Eng. Res. (IJEER) (2014)
3. Amar Babu, Y., Prasad, G.M.V.: An area and power efficient on chip communication architectures for image encryption and decryption. Int. J. Res. Eng. Technol. (IJRET) **3**(5) (2014)
4. Amar Babu, Y., Prasad, G.M.V.: Performance analysis and implementation of modified SDM based NoC for MPSoC on Spartan6 FPGA. Int. J. Res. Eng. Technol. (IJRTE) (2016)
5. Benini, L., De Micheli, G.: Networks on chips: a new SoC paradigm. Computer **35**(1), 70–78 (2002)

6. Bjerregaard, T., Sparso, J.: A router architecture for connection oriented service guarantees in the MANGO clockless network-on-chip. In: Proceedings of the Design, Automation and Test in Europe, DATE'05, pp. 1226–1231 (2005)
7. Dally, W.J., Towles, B.: Route packets, not wires: on-chip interconnection networks. In: ACM/IEEE Design Automation Conference (DAC), June 2001
8. Havemann, R.H., Hutchby, J.A.: High performance interconnects: an integration overview. Proc. IEEE **89**(5) (2001)
9. Hemani, A., Jantsch, A., Kumar, S., Postula, A., Oberg, J., Millberg, M., Lindqvist, D.: Network on chip: an architecture for billion transistor era. In: IEEE NorChip Conference, Nov. 2000
10. ITRS (International Technology Roadmap for Semiconductors: Semiconductor Industry Association), Dec 2015

A High-Speed Booth Multiplier Based on Redundant Binary Algorithm

Ranjan Kumar Barik, Ashish Panda and Manoranjan Pradhan

Abstract This article presents a high-speed Booth multiplier using a redundant binary algorithm which replaces the final addition stage. The redundant binary algorithm converts redundant binary to natural binary in constant time to provide a faster result. The Xilinx ISE design software 14.1 is used for the synthesis of the proposed architecture and implementation is done on Virtex-4 vlx15sf363-12 device for comparison purpose. The proposed architecture proves to be almost 74% faster than Booth multiplier using carry propagate adder (CPA), almost 65% faster than Booth multiplier based on carrying-lookahead adder (CLA), and more than 50% faster than vedic squaring architectures present in literature.

Keywords Redundant binary · Computer arithmetic · Vedic mathematics FPGA · Booth multiplier

1 Introduction

Array multiplication and Booth multiplication are the two most used multiplication techniques for digital hardware [1]. A signed binary multiplication technique commonly Booth multiplier [2] reduces the partial product rows count by a factor of N for radix 2^N encoding where N = 1, 2, 3…. The value of N decides the number of partial products and delay associated with the generation of partial products (PPs). Although higher value of N generates a lesser number of partial products, the delay associated with it is much higher. In natural binary, radix-4 Booth encoder is suitable for the minimization of partial products by half with lower complexity.

R. K. Barik (✉) · A. Panda · M. Pradhan
VSS University of Technology, Sambalpur, Burla, Odisha, India
e-mail: irkbarik@gmail.com

A. Panda
e-mail: ashish.panda14@gmail.com

M. Pradhan
e-mail: manoranjan66@rediffmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_56

This reduction in the number of PPs leads to the improvement in the computation of multiplier. Then generated partial products can be accumulated in various architectures such as array structure, redundant binary adder tree [3] or Wallace tree method [4]. In the Wallace tree method, the final result is the addition operation of sum row and carry row by using a carry propagate adder (CPA) or carry look ahead (CLA) adder. The carry propagation associated with addition operation is the major area of concern for designing a fast adder. The problem of carrying propagation can be solved by eliminating or limiting carry propagation within a small number of bits. In the redundancy representation [4, 5], some numbers are represented in multiple encoding which leads to limiting carry propagation to only one position.

In the paper Harata et al. [3], Authors have suggested a new computational rule for binary addition avoiding carry propagation. They claim to carry-free addition is possible by generating intermediate sum and intermediate carry. The authors in Phatak et al. [6], have presented a detailed analytical study of constant time addition and simultaneous format conversion considering fully as well as partially redundant representation.

The conventional RB number to NB number converter is designed with a CLA [3]. The RB number to NB number conversion using CLA is contradictory to the fact that RB number system is used to avoid the carry propagation where we require carry propagation to obtain the final result. So the design of efficient RB of NB converter other than CLA has been a continuous topic of development for the researcher. In 1987 [7], authors have presented an RB to NB converter using digit-by-digit generation, claiming application to non-restoring division, square root algorithms, and online algorithms. In the paper [8], authors have designed an RB to NB converter using both serial and look ahead modes. In the paper Sahoo et al. [9], a circuit level approach is reported to implement the equivalent bit conversion algorithm by Kim et al. [10, 11] for RB to NB conversion. A circuit based on predictable carry out technique was designed by the authors where carry out is independent of carry in for all values excluding initial condition. They claim improvement in delay and power over the conventional method of the converter, i.e., CLA. In their extended work [12], authors have modified their previously proposed circuit [9], utilizing the modified equivalent binary conversion algorithm (MEBCA). The authors have proposed a fast final adder considering 54×54 bits parallel multiplier and claim their final adder stage has 17% speed improvement over carry look ahead. Considering the result comparison of RB to NB converter with CLA the authors have claimed improvement to the proposed 54×54 bit multiplier architecture. This suggested multiplier architecture in [12], is considered to be novel using advantages of both NB and RB numbers system. However, the authors have also stated that the addition of subtract bit would not degrade the performance of the compressor stage only if the number of partial products plus one is divisible by four in radix 4 encoding system. To this fact, their architecture is not suitable for the multiplication where the number of generating PPs are even for radix-4 encoding, i.e., for 16×16 or 32×32 bit multipliers.

In this article, we have suggested a 16×16 bits multiplication architecture using both Booth algorithm and redundant binary algorithm. In the proposed

multiplier, the RB result is obtained without upsetting to the compression stage of the multiplier and the RB to NB conversion is performed using the HDL code considering FPGA platform. In recent years, Vedic mathematics [13] has emerged as a very effective approach to implementing efficient digital arithmetic operations. The vedic mathematics is important as it scales down the large calculations of the conventional mathematics to a simplified one. Considering these facts, the proposed multiplier architecture is also compared with recently suggested Vedic squaring architectures [14, 15]. The organization of the paper is as follows: Sect. 1 is the introduction. The proposed high-speed multiplication architecture using RB number system is discussed in Sect. 2. The comparison results of proposed architecture with various existing architectures are presented in Sect. 3. The conclusion to this article is presented in Sect. 4.

2 Proposed Multiplier Architecture Using RB Number System

The condition of redundancy representation is described [7] for any set [−α, β] of r or more consecutive integers that include 0 of the digit set in radix r. If r digit values are used, it offers a unique representation of each value within its range and irredundant number system. On the other hand, if more than r digit values are used and some values have multiple representations then it offers redundant number system. Considering digits sets [−1, 1] within radix 2 number system, the three digits value can have many encoding schemes with at least two bits. The value of n-bit of RB integer $m = [m_{n-1} m_n \ldots m_0]$ can be found as $\sum_{n=0}^{n-1} m_i * 2^n$. The RB digit ($r_i$) can be expressed by two NB digits (s_i, c_i) as $r_i = s_i - \overline{c_{i-1}}$.

Let A and B be the two NB numbers,

$$A + B = A - (-B) = A - (\bar{B} + 1) = (A - \bar{B}) - 1 = (A, \bar{B}) - 1 \quad (1)$$

Equation 1 shows any two NB numbers whose addition can be represented by redundant binary form $(A, \bar{B})$ with an additional subtract bit (-1). In the proposed multiplier architecture the RB coding scheme is employed in the concluding stage of addition followed by three steps RB to NB converter [10, 11]. In our architecture, the addition of subtract bit is performed outside of the Wallace tree because of this the performance of Wallace tree is not disgraced.

The addition of all partial products in a Wallace tree generates sum row (ws) and carry row (wc). The final sum (s) and final carry (c) are obtained after addition of subtract bit with ws and wc. The same RB coding scheme (Table 1) is used for final sum (s) and shifted final carry (c) to obtain the multiplication result in RB form. The proposed multiplier using RB coding technique is shown in Fig. 1.

Table 1 RB coding scheme

s_i	c_{i-1}	Value of RB digit (r_i)	RB digit (r_i)
0	0	−1	$\bar{1}$
0	1	0	0
1	0	0	0
1	1	1	1

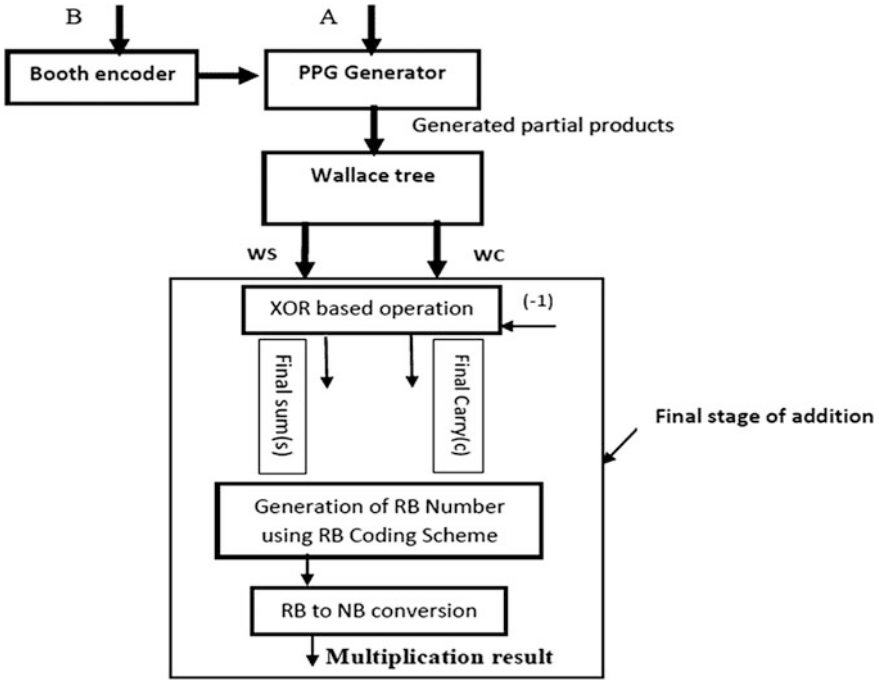
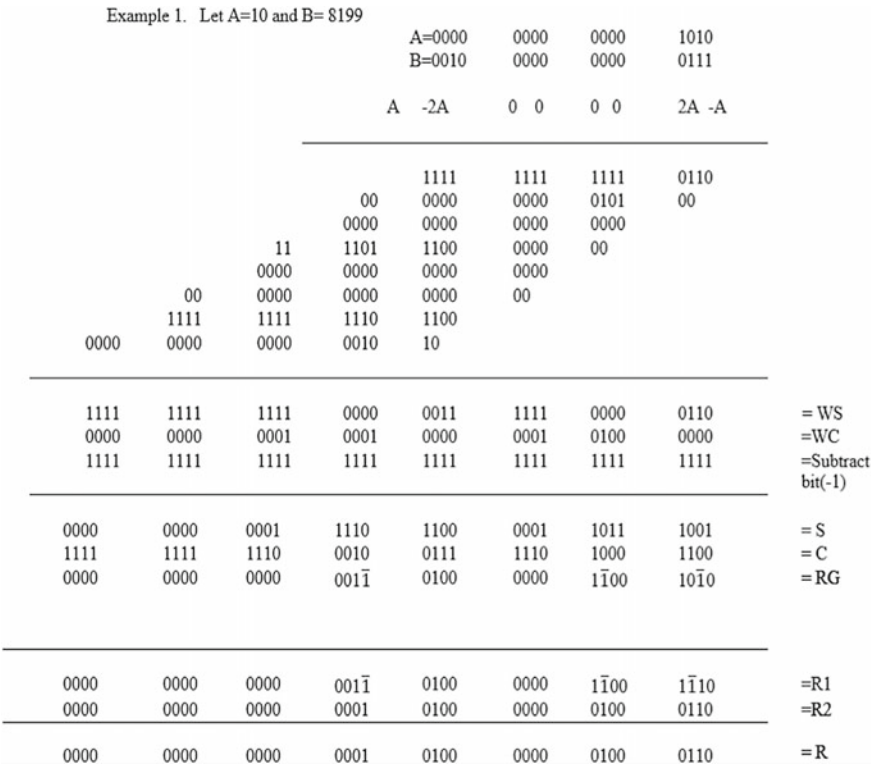


Fig. 1 Block diagram for proposed multiplication architecture

3 Results and Discussion

The whole 16 × 16 Booth multiplication architecture (as per Fig. 1) is also implemented (Virtex 4vlx15sf363-12 FPGA device) and simulated. The HDL (Hardware Description Language) program code for the proposed architecture is written in the design entry stage. The Logic synthesis and simulation is done using EDA (Electronic Design Automation) tool in Xilinx ISE 14.1 simulator. Table 2 shows delay comparison of the proposed architecture with architecture reported in [14, 15]. It is seen that proposed method achieves almost 72% reduction in delay as



WS= Sum row of Wallace tree
WC=Carry row of Wallace tree
S=Final sum row
C=Final shifted carry row
RG=Result in RB form (Grouping S_i and C_i)
R1= Result after stage 1
R2= Result after stage 2
R=Final result after discarding MSB

Fig. 2 Multiplication of two 16 bits number A = $(10)_{10}$ and B = $(8199)_{10}$ using RB to NB converter

compared to the reported technique in [14]. Our method has also more than 50% reduction in delay as compared to the squaring architecture reported by Sethi et al. [15]. Further, the proposed architecture proves to be almost 74% faster than Booth multiplier using RCA and almost 65% faster than Booth multiplier using CLA (Fig. 2).

Table 2 Delay comparison for proposed 16 × 16 multiplier architecture with [14, 15] (in ns)

Device vertex 4 vlx 15 sf363:-12	Booth's multiplier using RCA	Booth's multiplier using CLA	Vedic squaring unit reported in [14]	Square architecture reported in [15]	Proposed 16 × 16 multiplier architecture
Delay	36.65	27.43	33.39	18.56	9.2
4 input LUTs	880	629	294	233	662

4 Conclusion

This article presents a time-efficient design of Booth multiplier by using a redundant binary algorithm which replaces the final addition stage of the multiplier. The delay parameter is compared with modified Booth's multiplier as well as vedic squaring unit [14, 15]. The proposed architecture proves to be almost 74 and 65% faster than Booth multiplier using RCA and CLA. Further, our proposed multiplier has almost outperformed all other Vedic multiplier and Vedic squaring structures in terms of delay. This may be useful for high-performance microprocessor applications.

References

1. Goldberg, D.: Computer arithmetic. In: Patterson, D., Hennessy, J.L. (eds.) Computer Architecture: A Quantitative Approach. Morgan Kaufmann, Los Altos, California, Appendix A (1990)
2. Booth, A.D.: A signed binary multiplication technique. Q. J. Mech. Appl. Math. **4**, 236–240 (1951)
3. Harata, Y., Nakamura, Y., Nagase, H., Takigawa, M., Takagi, N.: A high-speed multiplier using a redundant binary adder tree. IEEE J. Solid-State Circ. **22**, 28–34 (1987)
4. Parhami, B.: Computer Arithmetic: Algorithms and Hardware Designs. Oxford University Press, Inc (2009)
5. Avizienis, A.: Signed-digit number representations for fast parallel arithmetic. IRE Trans. Electron. Comput. 389–400 (1961)
6. Phatak, D.S., Goff, T., Koren, I.: Constant-time addition and simultaneous format conversion based on redundant binary representations. IEEE Trans. Comput. **50**, 1267–1278 (2001)
7. Ercegovac, M.D., Lang, T.: On-the-fly conversion of redundant into conventional representations. IEEE Trans. Comput. **100**, 895–897 (1987)
8. Yen, S.-M., Laih, C.-S., Chen, C.-H., Lee, J.-Y.: An efficient redundant-binary number to binary number converter. IEEE J. Solid-State Circ. **27**, 109–112 (1992)
9. Sahoo, S., Gupta, A., Asati, A.R., Shekhar, C.: A novel redundant binary number to natural binary number converter. J. Signal Process. Syst. **59**, 297–307 (2010)
10. Kim, Y., Song, B.-S., Grosspietsch, J., Gillig, S.F.: A carry-free 54b × 54b multiplier using equivalent bit conversion algorithm. IEEE J. Solid-State Circ. **36**, 1538–1545 (2001)
11. Kim, Y., Song, B.-S., Grosspietsch, J., Gillig, S.: Correction to a carry-free 54 bx 54 b multiplier using equivalent bit conversion algorithm. IEEE J. Solid-State Circ. **38**, 159–159 (2003)

12. Sahoo, S.K., Shekhar, C.: A fast final adder for a 54-bit parallel multiplier for DSP application. Int. J. Electron. vol. **98**, 1625–1638 (2011)
13. Tirtha, S.B.K., Agrawala, V.S., Agrawala, V.: Vedic Mathematics, vol. 10. Motilal Banarsidass Publications (1992)
14. Kasliwal, P.S., Patil, B., Gautam, D.: Performance evaluation of squaring operation by Vedic mathematics. IETE J. Res. **57**, 39–41 (2011)
15. Sethi, K., Panda, R.: Multiplier less high-speed squaring circuit for binary numbers. Int. J. Electron. **102**, 433–443 (2015)

Evaluation of Channel Modeling Techniques for Indoor Power Line Communication

Shashidhar Kasthala and Prasanna Venkatesan G.K.D

Abstract Communication over existing electrical wiring has attracted many researchers as the last mile solution due to its reduced cost of installation. To make this medium as a feasible alternative to other communication medium, it is important to evaluate the performance of indoor electrical wiring. In this paper, an attempt is made to analyze and compare the various types of power line channel modeling techniques. The importance of modeling the cable parameters is also discussed. A sample residential power line network is considered and the channel frequency response and the channel capacity are compared for the various channel modeling techniques.

Keywords Channel capacity · Multipath propagation · Scattering parameters model · Transfer function · Transmission matrix model

1 Introduction

Recently, communication over indoor electrical wiring has drawn the attention of both industry and academia for its huge scope in network connectivity. In fact, it can be a viable alternative to wireless communication considering the deployment cost and the extent of power line network. However, for a full-scale commercial deployment, power line communication has a long way to go. This is primarily due to the challenges in understanding the critical channel parameter, viz., noise, multipath effect, and attenuation. Unfortunately, these three parameters vary with time, location and topology and thus making it further complicated [1–4].

S. Kasthala (✉)
Karpagam Academy of Higher Education, Coimbatore, India
e-mail: shashi_kb4u@rediffmail.com

Prasanna Venkatesan G.K.D
SNS College of Engineering, Coimbatore, India
e-mail: prasphd@gmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_57

To understand the behavior of these highly unpredictable power lines, extensive research has been carried out and various models have been developed. These models are primarily classified into top-down approach or bottom-top approach [5, 6].

The top-down approach of channel modeling is generally based on the multipath propagation model. In this model, the channel parameters are obtained largely by measurements. The developed models are easy to use but are of little flexibility. The bottom-top approach is an analytical method which requires a thorough knowledge of power line network. Though this approach is tedious, it gives the flexibility to model the power line in various scenarios [7].

These two models are achieved either in time or in the frequency domain. The choice in between these two modeling techniques is made based upon the size and complexity of the electrical network to be modeled [1]. Though adequate literature is found using both time-domain modeling and frequency domain modeling for top-down approach, it is not the case with bottom-top approach [8–10]. Frequency domain modeling is the preferred one for bottom-top approach [11–13]. In the recent past, even wavelet-based time-domain models are developed to evaluate the impact of loads on transfer functions [12].

In this paper, a sample residential electrical network is considered and a comparative study is carried out on the efficacy of top-down and bottom-top approaches. The parameters under investigation are transfer function and channel capacity. The model parameters obtained are in the context of Indian residential electrical wiring. The approximations usually made by researchers in modeling the cable parameters are also discussed. The power line noise is modeled using the Middleton Class A model and subsequently applied to estimate the channel capacity.

The work carried out is organized as follows. In the next section, the cable parameters are modeled. Based on this, the top-down approach emphasizing the multipath effect in time-domain modeling is discussed in Sect. 3 and in Sect. 4, the Bottom-top approach based on the TL model is discussed and is followed by a conclusion.

2 Model Parameters for Power Line Cable

To develop a channel model based on any approach, two parameters, viz., characteristic impedance (Z_O) and the propagation constant (γ) are required. These two basic parameters are obtained by the four intrinsic parameters viz. R, L, C and G measured in p.u. length [14, 15].

The residential electrical network in India usually contains 3 wires, the Live (L), Neutral (N) and Earth wire (PE) as shown in Fig. 1. These wires are loosely twisted together and can be either single core or stranded conductors with a cross-section of either 1.5 or 3 mm^2 based on the rating of electrical appliance used [16].

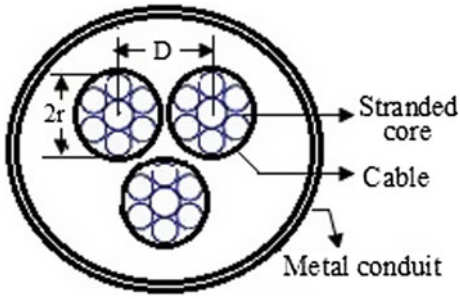


Fig. 1 Representation of 3-core residential electrical wiring

2.1 Resistance

The resistance of the single core cable is represented as

$$R = \frac{1}{\pi r \delta \sigma}\left[\frac{\frac{D}{2r}}{\sqrt{\left(\frac{D}{2r}\right)^2 - 1}}\right]. \tag{1}$$

The skin depth (δ) of the cable is given as

$$\delta = \frac{1}{\sqrt{\pi f \mu \sigma}}. \tag{2}$$

Here r is the conductor radius, D is the distance between conductors, σ is the conductivity and μ is permeability of cable. However, if the cable is stranded in nature, a correction factor (X_c) is required.

$$X_C = \frac{\left[\cos^{-1}\left(\frac{r_w - \delta}{r_w}\right) r_w^2 - (r_w - \delta)\sqrt{r_w^2 - (r_w - \delta)^2}\right]}{2 r_w \delta}. \tag{3}$$

where r_w is the radius of each strand in the conductor. The resistance obtained in (1) is for unit length of a single conductor. If a 2-wire transmission model is considered, the resistance has to be doubled. For a 3-wire residential network, the third wire has an influence on the overall resistance due to proximity effect which many of the researchers generally ignore.

2.2 Inductance

The inductance of a 2-wire cable is given as

$$L = \frac{\mu_r \mu_o}{\pi} \cosh^{-1}\left(\frac{D}{2r}\right) + \frac{R}{2\pi f}. \tag{4}$$

where μ_r is the relative permeability. For a stranded conductor, a correction factor a_C is to be considered, but many authors neglect this value [17]. The correction factor is given as:

$$a_c = \frac{n_w \pi r_w^2}{\pi r^2}. \tag{5}$$

The presence of the third conductor in the conduit will also affect the overall inductance leading to an additional factor of uncertainty.

2.3 Capacitance

Since the cables in residential electrical networks are placed in conduit pipes, the capacitive coupling effects of the conduit should also be considered along with the earth wire. The capacitance between any two conductors is given as:

$$C = \frac{\pi \varepsilon_o \varepsilon_r}{\ln\left[\left(\frac{D}{2r}\right) + \sqrt{\left(\frac{D}{2r}\right)^2 - 1}\right]}. \tag{6}$$

where ε_r is the relative permittivity of the cable. In residential power networks, the capacitance introduced between any of the conductor and conduit is neglected by many researchers since the difference in value is very minimal. But the conduit capacitance value can be significant in industrial or commercial power networks where the cable is surrounded by another electric field source.

2.4 Conductance

The conductance of the cable is given by the equation

$$G = 2\pi f C \tan\delta \quad S/m. \tag{7}$$

From (6), it can be understood that the approximation in the value of capacitance affects the conductance. As a usual practice, the distance and dielectric medium

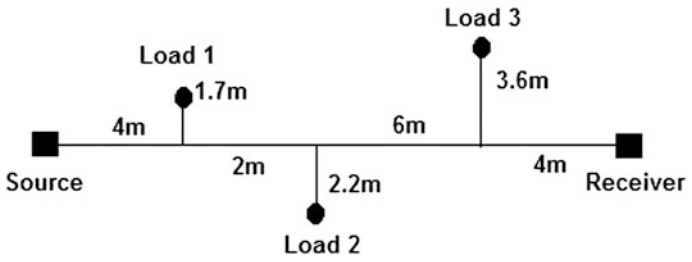


Fig. 2 Sample power line network

between the cables and between the cable and conduit is not maintained uniform, but due to the small cross-section of the conduit and the proximity of the cables, the impact of inhomogeneous can be neglected.

The four intrinsic parameters obtained from (1) to (7) are considered to determine the Z_o and γ. The characteristic impedance is given as

$$Z_o = \sqrt{\frac{(R+jwL)}{(G+jwC)}}. \tag{8}$$

and the propagation constant is given by

$$\gamma = \sqrt{(R+jwL)(G+jwC)}. \tag{9}$$

The sample network considered here is shown in Fig. 2. The cable used throughout the network is H07 V-U type with a similar cross-sectional area of 1.5 mm^2 and is connected with multiple domestic appliances. The cable is stranded in nature and has similar electrical properties throughout the network. The dimensions of the network considered are as per the recommendations in [16]. The source impedance considered for the network is 50 Ω and the load impedances are 61 Ω, 75 Ω and 45 Ω respectively. The line is terminated with 50 Ω.

3 Top-Down Approach

If the objective is to have a simple to use channel modeling technique with less number of parameters, then top-down approach is the better choice. One constraint of this approach is that the model developed for a specified network and specific frequency range that cannot be used for another network and different frequency bands.

Many researchers have contributed to this approach and various models have been developed so far accommodating different network topologies and frequency bands [10, 11, 18, 19]. The most preferred model in this approach is Zimmerman

and Dostert model developed on the concept of multipath propagation [9]. It is observed that the channel exhibits multipath nature because of the time-varying loads present in the network.

The model proposed by Philips [8] is based on the echo model which considers the power line has N signal flow paths. The transfer function for this model is given as

$$H(f) = \sum_{i=1}^{N} \rho_i e^{-j2\pi f \tau_i}. \tag{10}$$

where τ_i is the time delay of *i*th path and ρ_i is the result of transmission and reflection in that path.

The Philips echo model considers the transmission and reflections of the signal caused due to the impedance mismatches but does not take into account the attenuation of the signal. Also, the transfer function and impulse response estimated using this model are limited to constant length and constant load impedance.

The Zimmerman and Dostert [9] developed a model taking into account the signal attenuation. The CTF is expressed as

$$H(f) = \sum_{i=1}^{N} g_i e^{-\left(a_0 + a_1 f^k\right).d_i} e^{-j2\pi \frac{d_i}{v_p}}. \tag{11}$$

where N is the paths considered for the signal, g_i is weighing factor of the *i*th path, d_i path length, v_p is the cable phase velocity. The attenuation factor of the cable is obtained by the parameters a_o, a_1 and k. The weighing factor g_i is estimated by multiplying the transmission and reflection factors across the path.

For the network shown in Fig. 2, the weighing factor and the length of each path is obtained as mentioned in [9]. With the transfer function in (11), the amplitude response is as obtained in Fig. 3 and the capacity obtained from [20, 21] is as shown in Fig. 4 respectively.

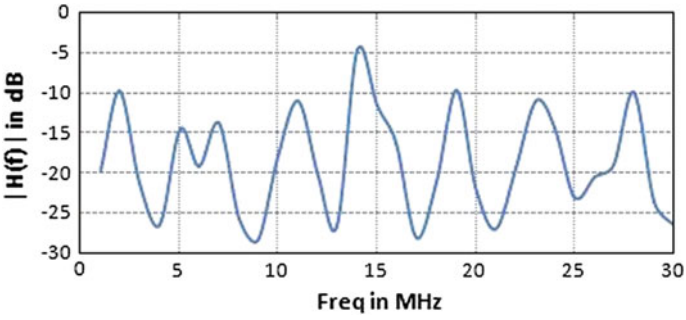


Fig. 3 Amplitude response using top-down approach

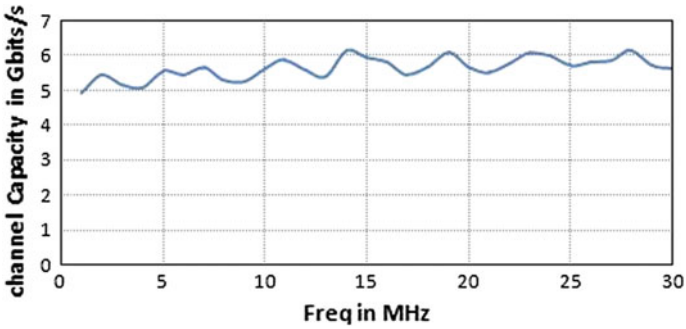


Fig. 4 Channel capacity obtained using top-down approach

Using (9) the amplitude, phase and impulse response of the electrical network line with less number of branches and constant load can be estimated easily. But the limitation the complexity that increases with the number of paths. This is true particularly in case of the residential power line.

4 Bottom-Top Approach

The bottom-top approach of channel modeling uses the TL theory to obtain the CTF. This approach is developed based on the complete information acquired from the electrical network, i.e., topology of the network, cables used and the loads present.

The various models based on bottom-top approach is obtained in literature. These models can be classified based upon transmission matrix [17, 22, 23, 24] or voltage ratio approach [12, 13, 25, 26, 27] or S-parameter matrix [1]. The limitation with the majority of these models is that they do not consider the practical aspects like grounding and coupling effects. This problem was addressed in [13, 27] by using the two-port networks.

4.1 Transmission Matrix Model

The transmission matrix model is also referred as ABCD model in few instances. In this model, the network is considered as a two-port networks and the relation between the voltage and current is obtained. Since the residential power line networks are complex in nature, the power line network can be segregated into N

single branch networks. The various useful parameters are obtained for each single branch. Once the matrices are achieved for individual networks as in [13, 24, 28] the matrix T of the entire network can be represented as.

$$T = \prod_{i=1}^{N} T_i = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}. \tag{12}$$

The transfer function which relates the matrix T and TL theory is expressed as

$$H(f) = \frac{Z_L}{T_{11}Z_L + T_{12} + T_{12}Z_S Z_L + T_{22}Z_S}. \tag{13}$$

4.2 S-parameter Model

Scattering parameter matrix also called as S-parameter matrix is the commonly used representation for transmission lines and loads [12]. The advantage of S-parameter is that it can be realized on a network consisting of different ampacity of cables. But the limitation of this model is that it cannot be dealt with nonlinear loads.

In this model also, the network can be divided into N networks to reduce the complexity. The scattering matrix for each network can be obtained and can be cascaded either by using chain scattering matrix method or signal flow graph method.

The S matrix obtained as mentioned in [1] is as follows:

$$\lfloor S \rfloor = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}. \tag{14}$$

The term S_{21} in (14) gives the channel transfer function. The S matrix and T matrix can be related as

$$\begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} = \begin{bmatrix} \frac{T_{21}}{T_{11}} & T_{22} - \frac{T_{21}T_{12}}{T_{11}} \\ \frac{1}{T_{11}} & -\frac{T_{12}}{T_{11}} \end{bmatrix}. \tag{15}$$

The amplitude response and frequency response of the transmission matrix model and S-parameter models are represented in the Figs. 5 and 6 respectively. It can be depicted from the figure that the amplitude response and the channel capacity for the both methods are in the same range with little variation. The variation is due to the methodology adopted and the parameters considered.

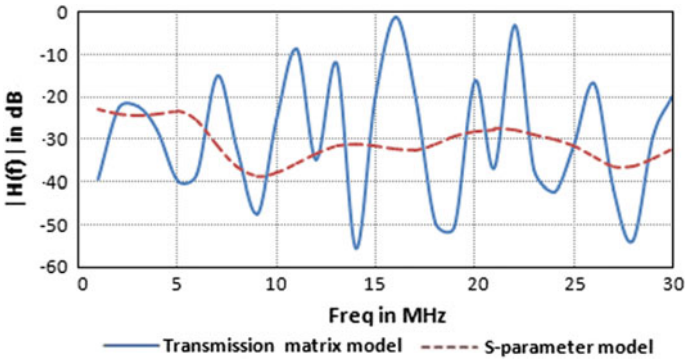


Fig. 5 Amplitude response using bottom-top approaches

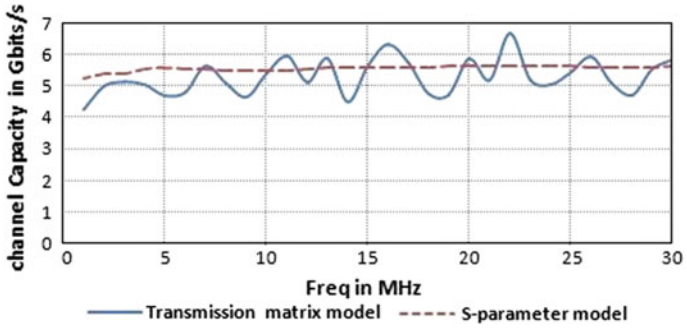


Fig. 6 Channel capacity obtained using bottom-top approaches

5 Conclusion

In this paper, various types of channel modeling techniques are analyzed. The choice of an approach is based upon the complexity and the specified application since each approach considers only certain key parameters. In this paper, it is also understood, that approximation of model parameters has an impact on the power line channel model. Many researchers have preferred bottom-top approach, in spite of its high computational efforts, because it gives a straightforward relation between the transmission line parameters and power line channel performance. However, there is still a need to develop a more accurate channel model taking into account the practical issues like radio interferences, weather interferences, etc.

References

1. Meng, H., Chen, S., Guan, Y.L., Law, C.L., So, P.L., Gunawan, E., Lie, T.T.: Modeling of transfer characteristics for the broadband power line communication channel. IEEE Trans. Power Deliv. **19**(3), 1057–1064 (2004)
2. Di Bert, L., Caldera, P., Schwingshackl, D., Tonello, A.M.: On noise modeling for power line communications. In: IEEE International Symposium on Power Line Communications and Applications (2011)
3. Anatory, J., Theethayi, N., Thottappillil, R.: Effects of multipath on OFDM systems for indoor broadband power-line communication networks. IEEE Trans. Power Deliv. **24**(3) (2009)
4. Maenou, T.: Study on signal attenuation characteristics in power line communications. In: IEEE International Symposium on Power Line Communications and Applications (2006)
5. Tonello, A.M., Versolatto, F.: New results on top-down and bottom-top statistical PLC modeling. In: 3rd Workshop on Power Line Communications (2009)
6. Zattar, H., Salek, L., Carrijo, G.: An evaluation of power line communication channel modeling for indoor environment application
7. Xu, W., Zhu, X., Lim, E., Huang, Y.: State-of-art power line communication channel modeling. Proc. Comput. Sci. **17**, 563–570 (2013)
8. Philips, H.: Modeling of power line communication channels. In: 3rd International Symposium on Power Line Communications and Applications (1999)
9. Zimmerman, M., Dostert, K.: A multipath model for power line channel. IEEE Trans. Commun. **50**(4), 553–559 (2002)
10. Galli, S.: A novel approach to the statistical modeling of wireless channels. IEEE Trans. Commun. **59**(5), 1332–1345 (2011)
11. Tilch, M., Zeddam, A., Moulin, F., Gauthier, F.: Indoor power line communications channel characterization up to 100 MHz part I: One parametric deterministic mode. IEEE Trans. Power Deliv. **23**(3), 1392–1401 (2008)
12. Barmada, S., Musolino, A., Raugi, M.: Innovative model for time-varying power line communication channel response evaluation. IEEE J. Sel. Areas Commun. **24**(7) (2006)
13. Galli, Stefano, Banwell, T.: A novel approach to the modeling of the Indoor power line channel—part II: Transfer function and its properties. IEEE Trans. Power Deliv. **20**(3), 1869–1878 (2005)
14. Andreou, G.T., labridis, D.P.: Electrical parameters of Low-Voltage Power Distribution cables used for power-line communications. IEEE Trans Power Deliv. **22**(2), 879–886 (2007)
15. Wagenaars, P., Wouters, P.A.A.F., van der Wielen, P.C.J.M., Steennis, E.F.: Measurement of transmission line parameters of three core power cables with common earth screen. IET Sci. Meas. Technol. (2009)
16. Carcelle, X.: Power Line Communications in Practice. Artech House (2009)
17. Anastasiadou, D., Antonakopoulos, T.: An experimental setup for characterizing the residential power grid variable behavior. In: 6th International Symposium on Power Line Communications and its Applications (2002)
18. Mlynek, P., Koutny, M., Misurec, J.: Multipath channel models of power lines. Elektorevue **1** (2) (2010)
19. Guzelgoz, S., Celebi, H.B., Arslan, H.: Statistical Characterization of paths in Multipath PLC Channels. IEEE Trans. Power Delivery **26**(1), 181–187 (2011)
20. Holter, B.: On the Capacity of MIMO Channel-A Tutorial Introduction. Norwegian University of Science & Technology
21. Tlich, M.: PLC channel characterization and modeling: OMEGA. European Union Project Deliverable D3.2 v.1.2 IST Integrated Project No ICT-213311 (2011)
22. Anatory, J., Theethayi, N.: Broadband Power Line Communication Systems: Theory & Applications. WIT Press, Southampton, Boston

23. Tonello, A.M., Zheng, T.: Bottom-up transfer function generator for broadband PLC statistical channel modeling. In: IEEE International Symposium on Power Line Communications and Applications, pp. 7–12 (2009)
24. Zheng, Y., Luo, G., Zhang, B., He, Y.: Research on power line as communication channel with multi-tap and multi-branch configuration. J. Netw. **8**(12) (2013)
25. Paul, C.R.: Analysis of Multiconductor Transmission Lines, pp. 46–62. Wiley (2004)
26. Versolatto, F., Tonello, A.M.: An MTL theory approach for the simulation of MIMO power line communication channels. IEEE Trans. Power Deliv. **26**(3), 1710–1717 (2011)
27. Banwell, Thomas, Galli, S.: A novel approach to the modeling of the Indoor power line channel—part I: Circuit analysis and companion model. IEEE Trans. Power Deliv. **20**(3), 655–663 (2005)
28. Kasthala, S., Prasanna Venkatesan, G.K.D.: Estimation of MIMO power line communication channel capacity using multi-conductor transmission line theory. In: 2nd International Conference on Applied and Theoretical Computing and Communication Technology (2016)

Power Analysis and Implementation of Low-Power Design for Test Architecture for UltraSPARC Chip Multiprocessor

John Bedford Solomon, D Jackuline Moni and Y. Amar Babu

Abstract Low-power architectures keeping in mind scalability presents a challenge to modern System on Chip and Network on Chip Designs. Especially, more so if these designs incorporate a Design for Testability Architecture too. DFT has become a De facto. From a Low-Power Scenario, it might seem easy to suggest a power down or power gating or clock gating or DVFS for a particular core to achieve this. But from a DFT perspective this presents a unique challenge as the scan chains and their allied clocks have to be active for verification to take place. Because if the power gated or clock gated low-power strategies can present difficulties to On-Chip Debug especially in modern SoC and NoC which tend to have long Test Data registers. The Drive to Low-Power Design should not impact yield or design confidence or test confidence. In this paper, a novel architecture is proposed to improve observability and controllability at individual core level while optimizing 20% of power consumption on UltraSPARC chip multiprocessor.

Keywords DFT · System on chip · Network on chip · Low power

1 Introduction

In SOC Design, Low-Power Design is a Key Design Issue [1–3]. Twice as much power is consumed during Testing than during Functional Operation; the reason being functional mode inputs are correlated as against Test Vectors which are

J. B. Solomon (✉) · D. J. Moni
School of Electrical Engineering, Center for Excellence in VLSI and Nanoelectronics, Karunya University, Coimbatore 641114, Tamil Nadu, India
e-mail: bedford.solomon@gmail.com

D. J. Moni
e-mail: moni@karunya.edu

Y. Amar Babu
LBRCE, Mylavaram, AP, India
e-mail: amarbabuy77@gmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_58

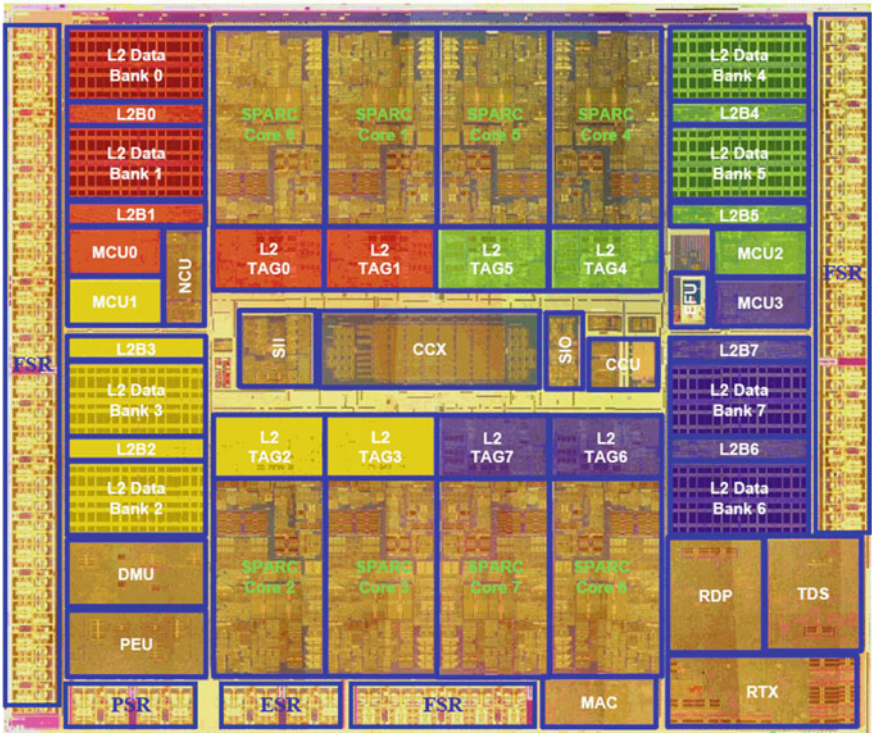


Fig. 1 UltraSPARC T2 CMP

Statistically independent [4]. Power is dissipated in traditional Design for Testability mechanisms in both the sequential scan cells and combinational logic [5, 6]. For the scan value to ripple down the combinational cone, redundant switching occurs which accounts for around 78% of the total test energy [7]. Figure 1, shows complex UltraSPARC T2 CMP which consists of 8 cores, each core has two integer ALU and one floating point unit. Each core support 8 threads and the overall architecture support 64 threads which can run concurrently. The T2 processor can run at 1.6 GHz speed. T2 processors support 4 MB L2 cache to support 64 concurrent threads which include eight times faster PCI express port. In this complex CMP architecture, Design for Test occupies the huge area and consumes power 8 times when compared with the previous T1 processor. In this paper a novel dynamic low-power design for the test has been proposed for UltraSPARC Chip multiprocessor T2 which eliminates multiple test register at the individual core level, individual L2 cache memory level thereby it optimizes power and area at the individual core level, cache level, and top-level architectures. The next section gives details about the proposed architecture.

2 Low-Power Design Techniques

Power Reduction Techniques fall into two broad categories namely the Dynamic Power or Switching power reduction Techniques and the Static or Leakage power reduction Techniques. The issue of Power Optimization for Dynamic Power at different levels of design abstraction from circuit to architecture to the system has been well researched in the literature [6].

2.1 Dynamic Power/Switching Power/Active Power

The clock is the signal which toggles the most in a circuit. It is also, therefore, the source of major dynamic power consumption. Clock Gating provides an effective technique with low design overhead [8]. The sizing of logic gates to effectively reduce the Switching Capacitance is a key design technique used to reduce dynamic power as well as to speed up critical paths [9]. Multiple threshold voltages technique of using dual vt transistors and assigning low vt transistors in critical paths and high vt transistors on noncritical paths can help in reducing dynamic power by 20% and leakage power by 50% [10]. Also, Critical Paths can be assigned High V_{dd} as against paths on the noncritical paths which are assigned Low V_{dd}. Operand isolation prevents unnecessary circuit activity by isolation design overheads which prevent redundant computation. The Technique of reducing or increasing the frequency of operation as per need on the fly is called Dynamic Frequency Scaling. If the same is done with voltage it is called Dynamic Voltage Scaling. Gating the Supply using sleep transistors is another frequently used method especially in regular structures to effectively reduce leakage power consumption.

2.2 Leakage Power/Standby Power

Apart from using Dual V_{th}, proper selection of Input Vector Control can deal up to 35% saving in Leakage power. Adding an additional transistor to gate the V_{SS} or V_{DD} line during idle mode effectively saves leakage power. Supply gating using sleep transistors is a very effective method to control leakage on regular structures, like SRAM typically employed in modern SOC Architectures extensively.

3 The Impact of Low-Power Design Techniques on Design for Testability

All low-power techniques add design overheads which impact test time. All low-power designs are essentially a tradeoff between performance and power and area. A circuit designed with power intent inevitably adds extra delay. Therefore, this takes extra cycles for a scan value to ripple down in a logic cone which is optimized for power. The issue gets further complicated in a modern SOC/NOC which tend to have very long Test Data Registers with multiple clocks and power domains. Typical issues are the blockage of the scan ripples. Further, ATPG tools cannot completely determine when power domains are switched on and off. More so if this is intended at the pattern level. The test engineer must have the ability to test all aspects of the power management structures such as the level shifters, isolation cells, retention states and power switches. The Analog Nature of some of these components poses a challenge. The placement of codec blocks and configured scan chains also poses a challenge. They must be in the same power domain else additional level shifters and isolation cells may be required.

3.1 Modern Design for Testability

Scan Testing is power intensive, up to 10x more than peak levels of normal operating mode. This generally leads to yield loss on account of device damage and false failures. Further, delay sensitive tests are affected by power spikes in today's nanometer devices. This is where power-aware testing plays a key role. Too much power consumption causes a problem during shift or capture mode. Excessive Power Consumption causes IR drop delays that prevent data from shifting at the target frequency. Also localized hot spot damage the device. The solutions to these have till now been to reduce the scan frequency. But this has the side effect of increased test time. Thereby increasing the test cost.

3.2 Existing Techniques of Power Reduction During Testing

These can be categorized into two parts, one during shift cycles, the other during capture cycles. The Technique of Low-power fill which reduces flop switching during shift effectively reduces up to 50% in test power. Adding gating logic, usually adds to large combinational logic cones. To reduce logic switching is another key method to reduce power during the shift. The IR drop increases dramatically during capture cycles, this is the case if the pattern count is higher, the higher the logic transitions, the higher switching activity and more the power

consumption. This can directly damage the device. The tradeoff is always between the switching activity versus the patterns count that defines the optimal test coverage. Designing wider power rails to accommodate to higher current flows has been the traditional solution.

4 Dynamic Scan Low-Power DFT

The Modern SOC/NOC uses a lot of reusable components. Many of these components have their own DFT. The resulting system/platform is built across vendors. Certain Test patterns for these SOC are complex. The idea is to reduce ATE tester time. Figure 2 shows the proposed low-power design for test architecture for UltraSPARC Chip Multiprocessor.

Proposed dynamic scan low-power design for the test can be implemented for any chip multiprocessor. We have selected UltraSPARC CMP as test architecture to evaluate our novel DS-LDFT technique. Proposed DS-LDFT has been inserted at an individual core level and memory level by using Synopsys DFT compiler tool and Perl scripting.

5 Power Analysis

DS-LDFT architecture evaluated at the top level UltraSPARC Chip multiprocessor by sending minimum test data and using minimum test time. Each core evaluated separately for observability and controllability for improved testability. Table 1 shows power consumption analysis of DS-LDFT compared with standard (SD) architecture with low-power and DFT techniques, Clock gating (CG), Power gating (PG), scan chain (SC), scan chain with power and clock gating (PG, CG, SC). Our proposed architecture shows low power consumption with improved testability. Refer (Table 1)

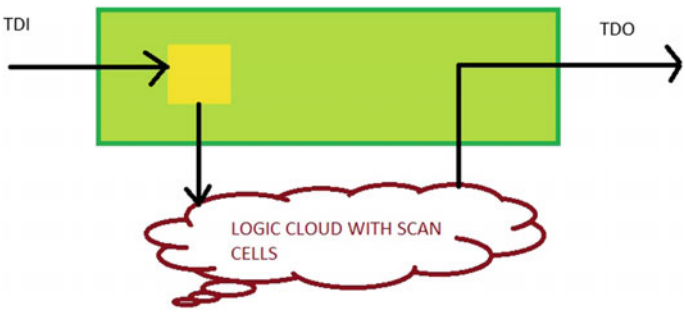


Fig. 2 Dynamic scan LDFT architecture

Table 1 Power analysis

Test architecture	SD (mw)	CG (mw)	PG (mw)	SC (mw)	PG + SC + CG (mw)	DS-LDFT (mw)
Memory control unit	4.89	2.78	2.98	5.8	2.6	2.34
Test control unit	0.29	0.19	1.78	0.35	0.15	0.14
Non cacheable unit	2.41	0.52	1.54	3.04	1.36	1.23
Entire chip	84.26	70.52	51.3	100.8	45.2681	40.7413

6 Conclusions

To improve the reliability of today's chip multiprocessors, we need to integrate SOC architecture with Low-Power Design and Design for Testability which is very challenging to optimize both power, area, and performance. But with dynamic scan Low-Power Design for Test Architecture we have shown on UltraSPARC CMP that power consumption can be get reduced by a factor of up to 20% while still maintaining high reliability of the chip. This work can be further enhanced with the network on the chip as test architecture.

References

1. Kowalczyk, A.: lst-generation MAJC dual microprocessor. ISSCC digest of technical papers (2001)
2. Zorian, Y., Marinissen, E.J., Dey, S.: Testing embedded core based system chips. In: International Test Conference, pp. 130–143
3. Marinissen, E.J., Zorian, Y.: Challenges in testing core based system ICs. IEEE communications magazine, pp. 104–109 (1999)
4. Varma, P., Bhatia, S.: A structured test re-use methodology for core-based system chips. In: International Test Conference, pp. 294–302
5. Marinissen, J.: A structured and scalable mechanism for test access to embedded reusable cores. In: Weaver, D., Germond, T. (eds.) International Test Conference, pp. 284–293. Prentice Hall (1994)
6. Roy, A.: A Novel cell-based heuristic method for leakage reduction in multimillion gate VLSI designs. In: 9th International Symposium on Quality Electronic Design ISQED 2008, (03/2008 Publication)
7. Levitt, M.E.: Testability, debuggability, and manufacturability of UltraSPARC?-I microprocessor. In: International Test Conference, pp. 157–166. International Technology Roadmap for Semiconductors (2015)
8. Roy, K.: Low-power design techniques and test implications. In: Power-Aware Testing and Test Strategies for Low Power Devices, (2009 Publication)
9. Hirech, M.: EDA Solution for poweraware design-for-test. In: Power-Aware Testing and Test Strategies for Low Power Devices, (2009 Publication)
10. Piguet, C.: Ultra-low-power signal processing in autonomous systems. In: Energy Autonomous Micro and Nano Systems Belleville/Energy Autonomous Micro and Nano Systems (2013)

Power Optimization for Arithmetic Components in Assistive Digital Devices

Mansi Jhamb and Gitanjali

Abstract The Wireless Body Area Network is a wireless sensor network that supports a wide range of novel assistive devices for healthcare and biomedical applications. The design goals of WBAN components are portability, low power dissipation and speedy operation. In this paper, a comparative analysis of domino logic-based arithmetic component has been carried out to analyze its performance with respect to WBAN applications. This component has less delay overhead (due to dynamic logic application in designing the functional block), minimal power consumption (a benefit of asynchronous design methodology being used), and has high throughput (due to pipelined processing).

Keywords WBANs-Wireless body area networks · DSPs-Digital signal processors · CD-Completion detector · FB-Functional block

1 Introduction

The WBANs [IEEE802.15.6] consist of miniature-devices such as sensors, transceivers, batteries, and embedded DSP processors which constitute together to form assistive digital devices [1–3]. Figure 1 depicts a typical WBAN for real-time healthcare systems. A gateway device (PDA) sends the collected information from sensor nodes onto the network, which is collected by the server and is sent to multiple health monitoring applications.

In DSPs, arithmetic components are the key elements. Hence, designing an area-delay-power efficient arithmetic component is the design criteria to achieve high performance. Asynchronous circuit designs are known for their low dynamic power dissipation, early computation completion sensing and high-speed operation [4, 5]. Since clock synchronization is not required in these circuits, they offer higher

M. Jhamb · Gitanjali (✉)
University School of Information, Communication and Technology, GGSIPU, Sector-16C, Dwarka, New Delhi, India
e-mail: gitu.2408@gmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_59

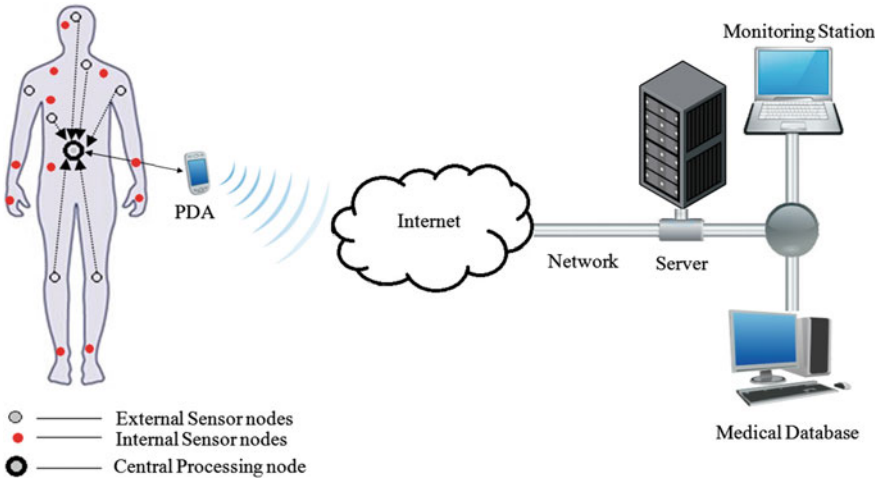
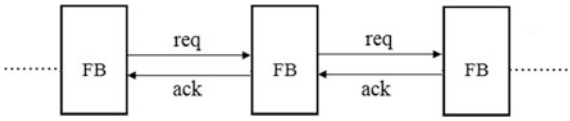


Fig. 1 Typical working of a WBAN for healthcare systems

Fig. 2 Asynchronous pipeline design



throughput and lower power consumption than their clocked counterparts [6]. In this paper, the arithmetic component is an asynchronous pipelined Domino logic-based Adder. Domino logic has been used for speedy operation [7].

2 Pipelining

To achieve high performance, digital systems use pipelining concept. It boosts system performance and provides high throughput via parallel processing [8].

Asynchronous pipelines: These are clock-less pipelines that use bidirectional communication (achieved by a handshaking protocol) for coordination among the functional blocks of a pipeline. Figure 2 shown depicts asynchronous pipeline functionality where, *req*—is the request signal to initiate computation procedure for FB and *ack*—acknowledgment received by the sender from the receiver.

Advantages of asynchronous pipelines:

(i) Multiple data items processing [8].
(ii) Automatic flow control is provided by asynchronous circuits,
(iii) Dynamic power consumption is low.

Dynamic logic-based pipelines: The classic design was proposed by Williams and Horowitz [9]. These pipelines use an implicit latching feature of dynamic circuits hence storage registers are not required in between functional stages. Asynchronous dynamic pipelines provide the following advantages:

(i) Latch elimination.
(ii) Critical path delay reduction.
(iii) Decreased power consumption.

2.1 The Classic Dynamic Pipeline Style—PS0 Pipeline

It is a dynamic logic-based latch-less pipeline proposed by Williams and Horowitz [9]. Figure 3 shown depicts a PS0 pipeline structure.

In this figure, *pcr*-acknowledgement signal from the successor block, *pcl*-acknowledgment signal going to the predecessor block, *CD*-Completion Detection circuit to generate acknowledgement and FB is a 1-bit dual rail domino Adder [10]

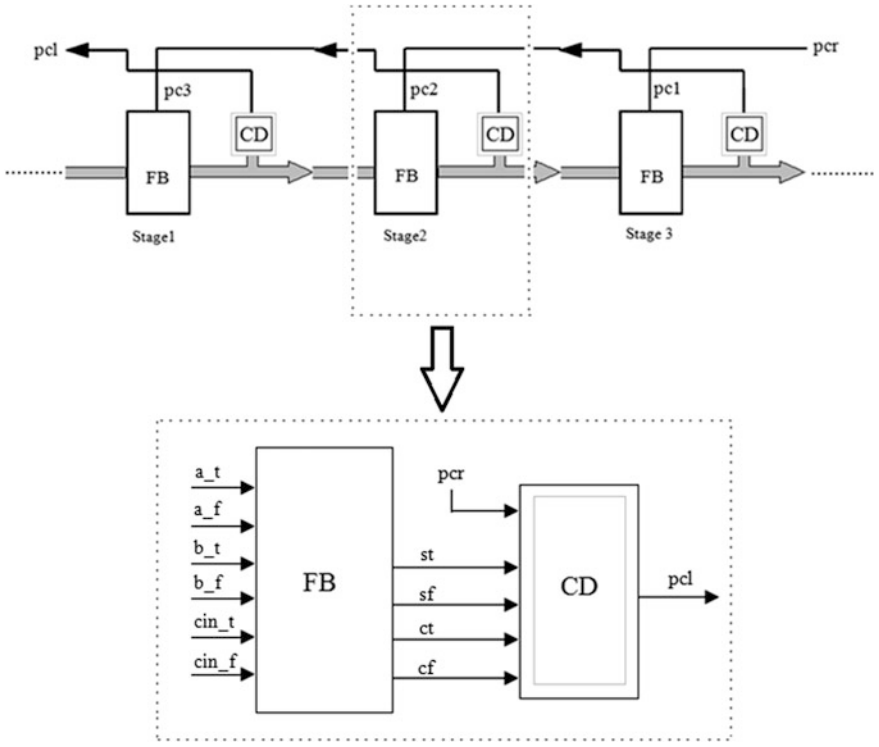


Fig. 3 PS0 pipeline functional design

in which, *a_t, a_f, b_t, b_f, cin_t, cin_f* are the dual rail inputs to the functional block, *st, sf, ct, cf*, are the dual rail outputs.

Working of PS0 Pipeline: (i) Stage 1 enters evaluation phase (ii) Stage 2 enters evaluation phase (iii) Stage 3 enters evaluation phase (iv) CD of stage 3 sends an acknowledgement pc1 indicating the completion of evaluation and hence initiates the precharge mechanism for stage 2 (v) Stage 2 gets precharged. (vi) CD of Stage 2 detects precharge completion and sends an evaluation enabling signal (pc2) to stage 1 so that stage 1 can start evaluating and then CD (stage 3) generates final acknowledgment *pcl.*

3 Implementation of Functional Block and Completion Detector

When the functional block completes its computation process, a circuit is required to detect the computation completion. Here, this has been done using a 2-bit completion detector circuit [11] which makes use of NOR gates to generate individual acknowledgment signals for dual rail outputs and all these signals are combined using a C-element [12, 13]. In this paper, the performance of pipelined adder has been analyzed in the presence of different C-elements in the completion detector circuit. Figure 4a shows the block diagram of the completion detection circuit used here. In this fig. **st**, **sf**, (sum) **ct**, **cf**, (carry) are the dual rail outputs and **pcn** where n = 1, 2, 3 … is the final acknowledgment generated for nth stage of the pipeline. The MOS-level circuit diagram is shown in Fig. 4b.

The C-element block shown in the diagram will be replaced by four different C-element designs proposed in the literature which are Martin's C-element [12], Van Berkel's C-element [14], Sutherland's C-element [13], Dynamic C-element [15].

Martin's C-element [12]

The Martin's C-element [12] has a weak inverter in feedback path at its output node. The feedback inverter should be designed to be weak enough so as to get overpowered by the PMOS PUN (pull-up network) and NMOS PDN (pull-down network), so that the circuit retains its previous output (Z) state when both the inputs X_1, X_2 change to different states {0, 1 or 1, 0}. Therefore optimum W/L ratios [13] have been chosen in order to avoid race condition at pcn_b node (depicted in Fig. 5a) [16]. This circuit is a semi static implementation in which set function has been implemented by the NMOS pull-down network and reset function has been implemented by the PMOS pull–up the network. Figure 5a shows the MOS-level circuit schematic of this C-element and the layout design for this C-element is depicted in Fig. 5b.

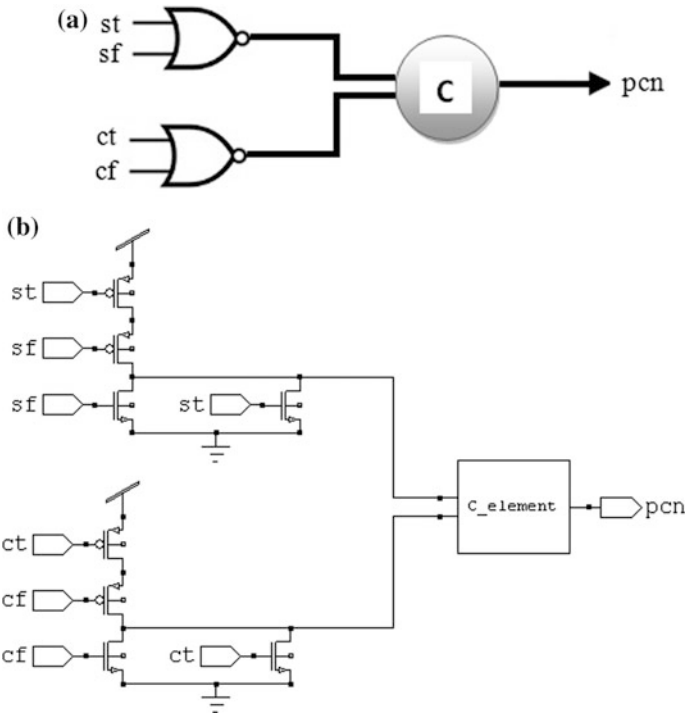


Fig. 4 **a** Block diagram of completion detector circuit [11] **b** MOS level circuit diagram of completion detector circuit

Sutherland's C-element [13]
The Sutherland's static implementation of C-element [13] also contains a set and reset circuitry, but the weak inverter feedback mechanism has been replaced by a second Pull up network to provide set' and pull down network to provide reset' functionality. It is a ratio-less circuit which can be generalized to three or more inputs. Figure 6a shows the MOS-level circuit schematic of this C-element and the layout design for this C-element is depicted in Fig. 6b.

Van Berkel's C-element [14]
Van Berkel's static implementation of C-element [14] also contains a set and reset circuitry. The difference lies in the feedback mechanism, that the feedback path consists of three transistors in pull-up network or pull-down network. Like Sutherland's C-element [13] it is also ratio-less. Another advantage is that it is symmetric with respect to inputs. The disadvantage is that only two input design exists for this C-element, thus it cannot be generalized. The design constraints are similar to that of Sutherland's implementation, i.e., feedback transistors should be minimum in size to maintain their state holding capability at the output node [13]. Therefore optimum W/L ratios have been chosen for all the transistors. Figure 7a

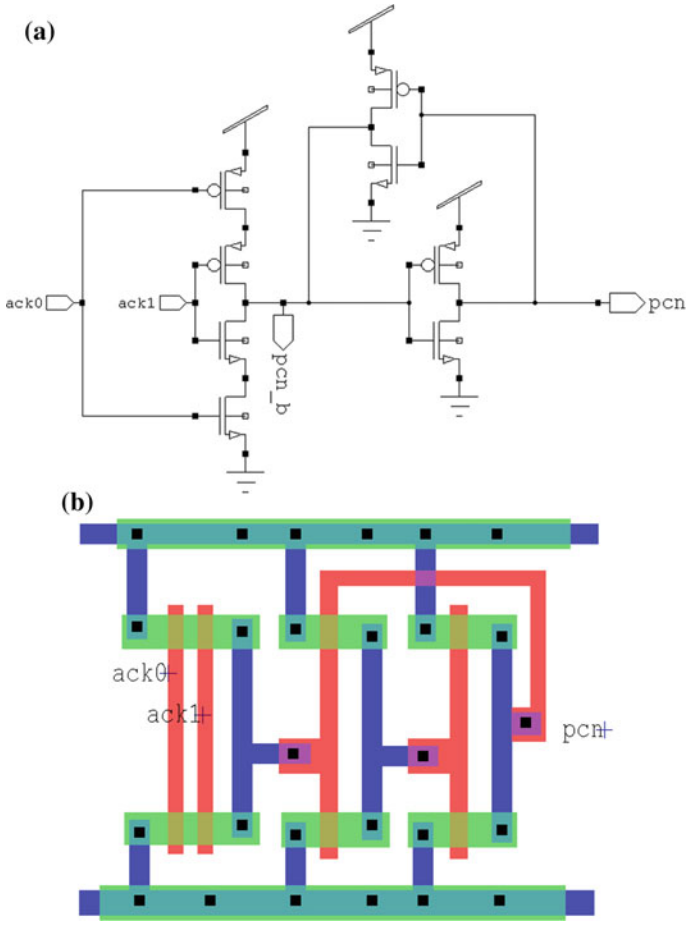


Fig. 5 **a** MOS-level schematic of Martin's C-element [12] **b** Layout design of Martin's C-element [12]

shows the MOS-level circuit schematic of this C-element and the layout design for this C-element is depicted in Fig. 7b.

Dynamic C-element [15]

The dynamic implementation of C-element [15] also contains the set and reset circuitry but the weak feedback inverter has been eliminated leading to high-speed operation. It is also a ratio-less circuit. Figure 8a shows the MOS-level circuit schematic of this C-element and the layout design for this C-element is depicted in Fig. 8b.

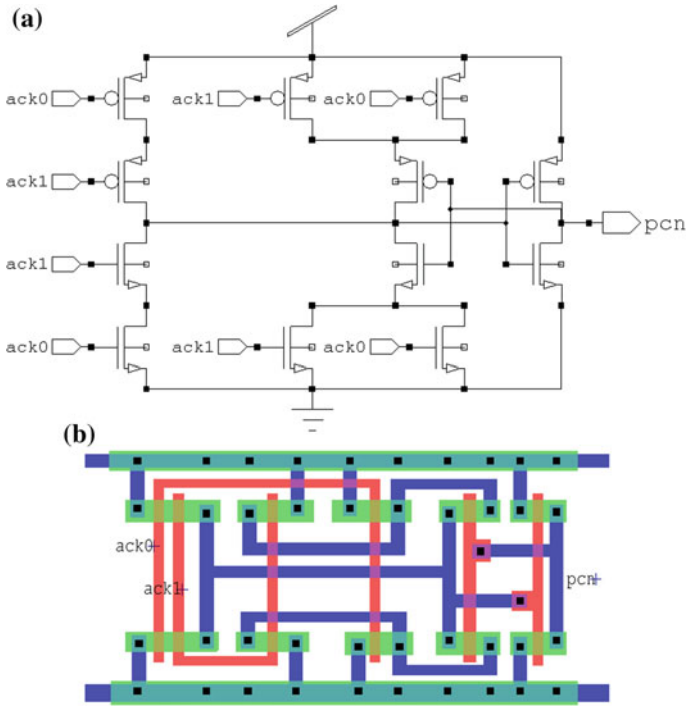


Fig. 6 **a** MOS-level schematic of Sutherland's C-element [13] **b** Layout design of Sutherland's C-element [13]

3.1 Functional Block

The functional block employed for examining the behavior of various completion sensing circuits in a dynamic pipelined structure is a 1-bit dual rail domino Full adder [10]. Domino CMOS circuits are an alternative to steady-state circuits in terms of high-speed operation, minimal area overhead and low latency [7, 10]. The inputs to this DUT are dual rail inputs (*a_t, a_f, b_t, b_f, cin_t, cin_f* (carry inputs)) resulting in dual rail outputs *st, sf* (sum), *ct, cf* (carry output) and *pc* is the precharge synchronizing signal. The MOS level circuit diagram of FB is given in Fig. 9.

4 Simulation and Results

Elaborate SPICE-level simulations were carried out for all the four pipelined designs using HSPICE (© AVANT Corporation!) at 90 nm TSMC CMOS technology with voltages varying from 0.8 to 1.2 V and temperature at 25 °C. For simplicity, completion detector circuits have been named as follows:

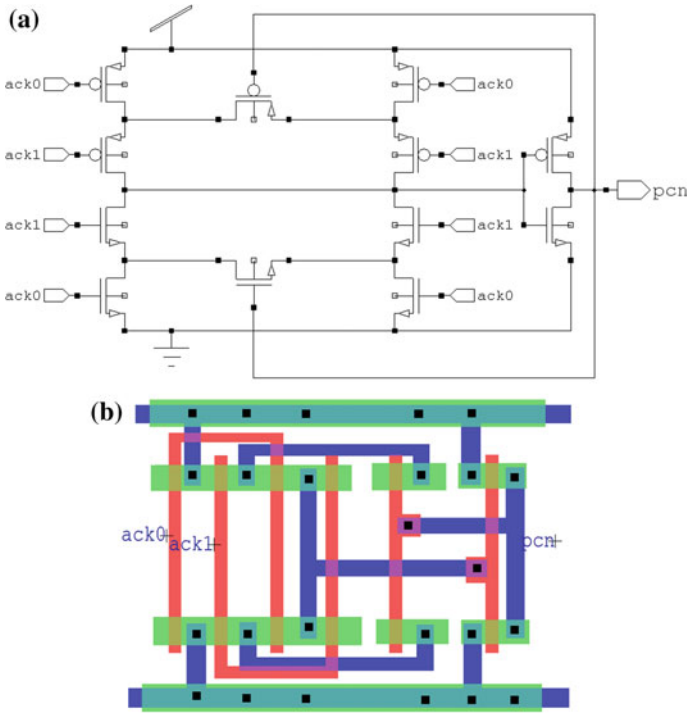


Fig. 7 **a** MOS level schematic of Van Berkel's C-element [14] **b** Layout design of Van Berkel's C-element [14]

CD1—completion detection circuit using Martin's C-element [12].
CD2—completion detection circuit using Sutherland's C-element [13].
CD3—completion detection circuit using Van Berkel's C-element [14].
CD4—completion detection circuit using Dynamic C-element [15].

Total Power Dissipation: The variation of power dissipation with supply voltage is shown in Fig. 10 which depicts that CD1 dissipates the highest power among all CDs. At vdd = 1.2 V, CD4 dissipates the lowest amount of power which is 88.32% lower than CD1. CD4 incurs 80.1% less power dissipation than CD1 at lowest voltage (vdd = 0.8 V).

Critical Delay Analysis: As depicted in Fig. 10, the best delays were obtained at vdd = 1.2 V, at which CD4 incurs the lowest delay (64% lower than CD1). At vdd = 0.8 V, CD1 incurs the largest delay which is 64.5% higher than CD4.

Energy Dissipation: The energy dissipation variation is depicted in Fig. 11 for all CDs. The highest energy is dissipated by CD1 at vdd = 1.2 V which is 95.8% higher than CD4. CD4 dissipates 95% lower energy than CD1 at vdd = 0.8 V.

Throughput and Latency: The cycle time for a PS0 Pipeline is:

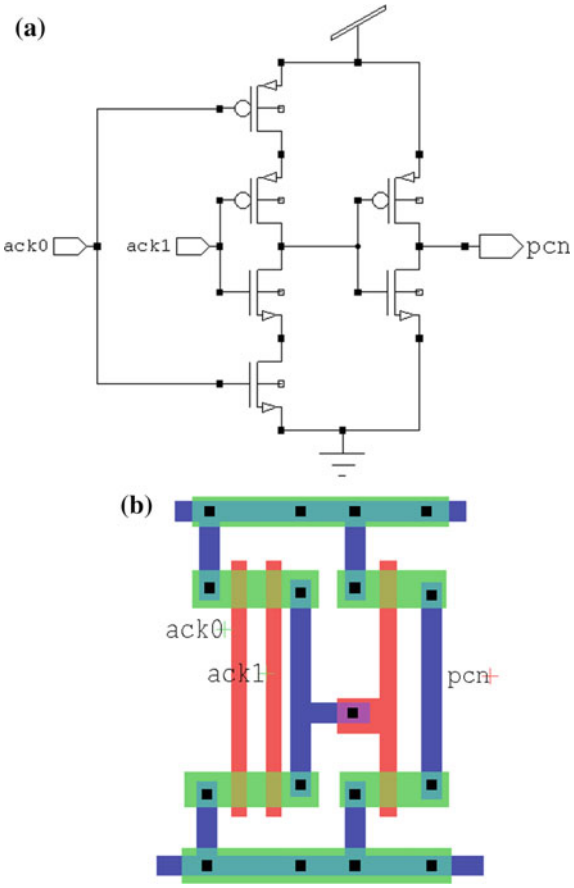


Fig. 8 **a** MOS level schematic of dynamic C-element [15] **b** Layout design of dynamic C-element [15]

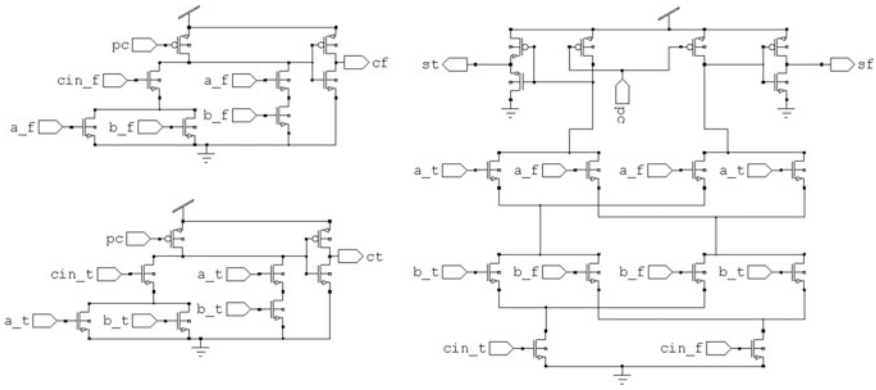


Fig. 9 1-bit dual rail domino full adder [10]

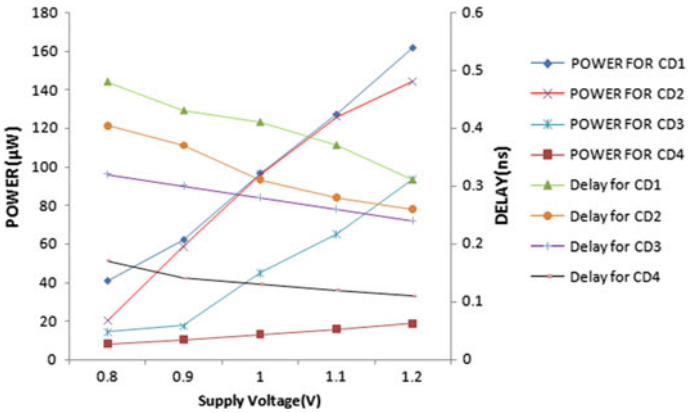
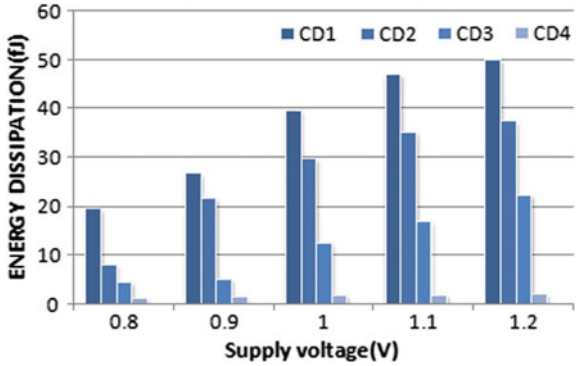


Fig. 10 Power and delay variation of CDs

Fig. 11 Energy dissipation



$$\mathbf{T_{Cycle} = 3T_{FB} + T_{PRE} + 2T_{CD}} \quad (1)$$

[8], where, $\mathbf{T_{FB}}$ = Evaluation time of functional block, $\mathbf{T_{PRE}}$ = Precharge time and $\mathbf{T_{CD}}$ = Time taken for acknowledgment generation by CD. The per-stage forward latency is $\mathbf{T_{FB}}$ [10]. CD4 gave the best throughput results at vdd = 1.2 V which was 117% higher than that using CD1. At vdd = 0.8 V, CD1 has 93.4% lower throughput than CD4. The throughput variation is depicted in Fig. 12 and latency variation with supply voltage is shown in Fig. 13.

Transistor Count: More the no. of transistors, higher the complexity of the circuit. As depicted from Fig. 14, the circuit complexity is highest for CD2 and CD3. The lowest transistor count has been obtained for CD4.

Layout Area: The VLSI Design layout has been designed using two-metal wire approach for each of the C-elements along with DRC and LVS-check verification.

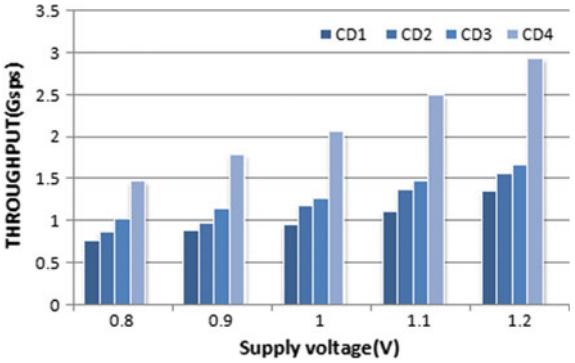


Fig. 12 Throughput variation with supply voltage

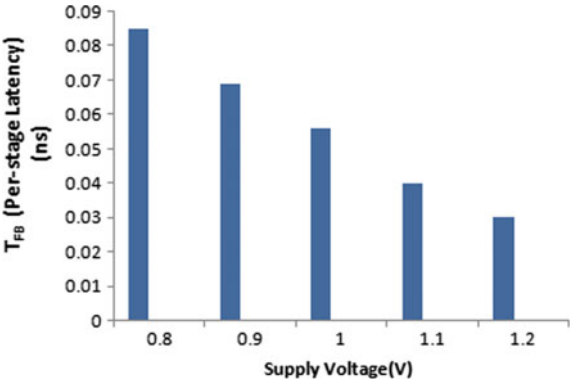


Fig. 13 Latency variation with supply voltage

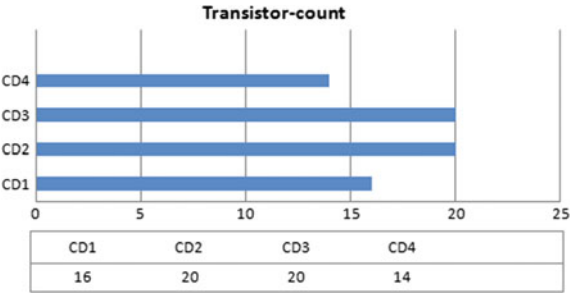


Fig. 14 Transistor count for all CDs

Figure 15 depicts the area required for different C-element implementations. Area requirement is highest for Sutherland's C-element [13] whereas dynamic C-element [15] requires the lowest area.

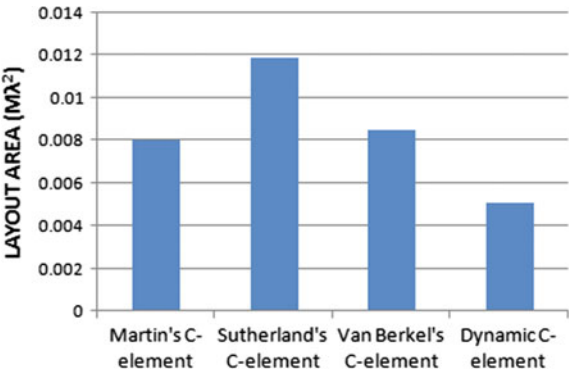


Fig. 15 Layout area comparison

5 Conclusion

The simulation results depict that asynchronous pipelined adder implementation using CD4 (with dynamic C-element) provides numerous benefits in terms of energy, power optimization, area, transistor count, delay and throughput depicting its ability to achieve high performance in WBAN applications.

References

1. Masuch, J., Delgado-Restituto, M.: Ultra low power transceiver for wireless body area networks. In: Analog Circuits and Signal Processing. Springer (2013). ISBN 978-3-319-00098-5
2. Yang, G.Z. (ed.): Body Sensor Networks. Computer Science, HCI. Springer (2014). ISBN 978-1-4471-6347-9
3. Kasun, M., Thotahewa, S., Redoute, J.-M., Yuce, M.R.: Ultra wideband body area networks. Engineering Circuits and Systems. Springer (2014). ISBN 978-3-319-05287-8
4. Brzozowski, J.A., Seger, J.H.: Asynchronous Circuits. Springer (1995). ISBN 978-1-4612-8698-1
5. Rahman M.Z., Kleeman, L., Habib, M.A.: Recursive approach to the design of parallel self-timed adder. IEEE Trans. VLSI Syst. (2014)
6. Birtwistle, G., Davis, A.: Asynchronous Digital Circuit Design. Springer (1995). ISBN 978-3-540-19901-l4
7. Kang, S.M., Leblebici, Y.: CMOS Digital Integrated Circuits. Tata Mc Graw-Hill (2003). ISBN 978-0-07-053077-5
8. Nowick, S.M., Singh, M.: High-performance asynchronous pipelines an overview. IEEE Des. Test Comput. **28**(5), 8–22 (2011)
9. Williams, T.E.: Self-timed rings and their application to Division. Ph.D. thesis, Computer Systems Lab, Stanford University (1991)
10. Weste, N., Harris, D.: CMOS VLSI Design: A Circuits and Systems Perspective. Addison Wesley (2004)
11. Xia, Z., Hariyama, M., Kameyama, M.: Asynchronous domino logic based pipeline design based on constructed critical data path. IEEE Trans. VLSI Syst. **23**(4) (2015)

12. Martin, A.J.: Formal progress transformation of VLSI circuit synthesis. In: Formal Development of Programs and Proofs, pp. 59–80. Addison Wesley (1989)
13. Sutherland, I.E.: Micropipelines Communications, vol. 32, pp. 720–73. ACM (1989)
14. Berkel, K.V.: Beware the isochronic fork. Integr. VLSI J. **13**(0.2), 103–128 (1992)
15. Cheng, F.C: Practical design and performance evaluation of completion detection circuits. In: IEEE International Conference in Computer Design (ICCD), pp. 354–359 (1998)
16. Shams, M., Ebergen, J., Elmasry, M.: Modelling and comparing CMOS implementation of C-element. IEEE Trans. VLSI Syst. **6**(4), 563–567 (1998)

EEG Artifact Detection Model: A Landmark-Based Approach

S. Mouneshachari, M. B. Sanjay Pande and B. N. Raveesh

Abstract Human Intelligent Assessment is one of the challenging task corresponding to engineering perspective. This could be made possible by combining multidisciplinary fields that is Psychology, Neurology, Engineering, and so on. This paper is one such attempt to find a new methodology for the efficient analysis of Human Intelligent Index using signal processing concepts by including compression and analytical techniques.

Keywords EEG artifact · Landmark-based · Discrete wavelet (daubechies) transform · Sphericity test

1 Introduction

Quotient assessment of intelligent index is one of the major trending topic of interest in psychology and medical sciences, currently, it has entered the stream of engineering as well [1, 2]. The perspective of engineering may be the solution applicable for effective and efficient analysis and proper estimation of the same. Since centuries the field of psychology is introducing one or the other scales in new dimension [3–6]. Intelligent Quotient was one of the major component played since centuries and took a right shape when it was properly declared by Thurstone [7], later, the improvements and developments were introduced by number of psychologists and scientists [8].

S. Mouneshachari (✉) · M. B. Sanjay Pande
CERSSE, Jain University, 52, Bellary Road, Hebbal, Bengaluru, India
e-mail: mounesh_s@yahoo.co.in

M. B. Sanjay Pande
e-mail: rakroop99@gmail.com

S. Mouneshachari · M. B. Sanjay Pande
Department of Computer Science and Engineering, GMIT, Davangere, India

B. N. Raveesh
Department of Psychiatry, MMCRI, Mysore, India
e-mail: raveesh9@gmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_60

The rigorous studies and exercises were made by Bar-on, Goleman, and others to think about another component of Intelligence Index and they have called it as Emotional Quotient. Mouneshachari S et al introduced a proper arrangement of these scales in a different proportion for the estimation of Intelligent Index and called it as S-Quotient [9]. This paper deals with the detection of EEG artifact for an efficient estimation of item assessment of any scale.

2 Related Work

2.1 EEG Artifact

Electroencephalography (EEG) is a captured/recorded electrical activity of the brain. It was discovered by a great German scientist, psychiatry Hans Berger in the year 1924 [10]. Brain–Computer Interface applications require one or the other input resources to fetch the brain activities. Hence, EEG is one of the easily available input resources for BCI applications [11].

2.2 Discrete Wavelet Transform

DWT mainly deals with the compression of the data using mathematical interpretations. Figure 1 shows the approximation and detail coefficients of f(t) based on certain conditions [12]. Articles [12, 13] describes the mathematical model used to find f(t) and its related approximation and detail coefficients.

2.3 Sphericity Estimation

The ratio between geometric mean and the arithmetic mean of principal axes of the ellipse [14] is called as sphericity and is given by the Eq. 1

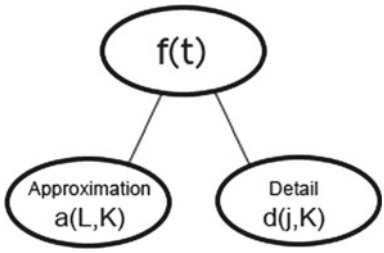


Fig. 1 Discrete wavelet transform

$$sphericity = \frac{2\sqrt{d1 + d2}}{d1 + d2} \tag{1}$$

where d1 and d2 are the lengths of principal axes.

Affine transformation represents the triangular transformations and it can be used as an alternative model for the estimation of sphericity [15]. Affine transformation does the mapping from x to u [14], where x and u belongs to R^2 [14]. Sphericity test has a number of applications, specifically in signal detection [16]. The sphericity value indicates the similarity of two triangles i.e., sphericity value 1 indicates both triangles are similar and smaller the value of sphericity, smaller the similarity.

2.4 Software Tools Used

Matlab is one of the most widely used software tool in the field of research and developments [17], essentially required for research in engineering [18]. ThinkGear [19] is a software driver used to access NeuroSky device. Neurosky communicates JSON objects to transfer the data from headset containing ThinkGear to the Client machine.

3 Methodology

Figure 2 depicts the methodology of the proposed work.

3.1 Capture Single Channel EEG Signal

A person is referred as a subject of interest to capture EEG signal in the presence of the expert. The precaution of informing about the need of assessment has been instructed. One of the Item from IQ scale will be provided to the subject. Higher attention and hyperactivity of the brain is required during the recording of the EEG

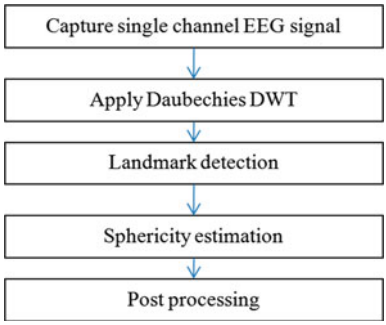


Fig. 2 Methodology for the EEG artifact detection

signals. Hence, the proper training and environment have been provided to the subject to attain around 80–90% of the attention. Each recording was conducted for around 10–12 s at the rate of 512 samples/seconds then stored in the .dat format.

3.2 Discrete Wavelet (Daubechies) Transform

Efficient approaches may require compression techniques to reduce the time and space complexities. One such flexible compression technique is Discrete Wavelet Transform. At each level of compression i the approximation a_i is separated into approximation a_{i+1} and detail d_{i+1} coefficients.

3.3 Landmark Detection

Landmarks are the local properties of the signal. Landmarks detection in the digitally sampled signal is quite simple which is as shown in Fig. 3. The next highest +ve value for the initial point of the signal is the point at 2, then next lowest –ve value for point 2 is the point at 3. Similarly from 3 to 4 and so on, and these 1, 2, 3, 4, and so on are regarded as landmarks of the signal.

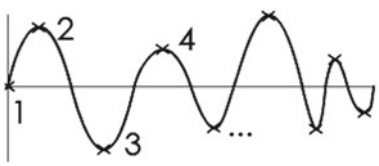


Fig. 3 Landmarks in signal

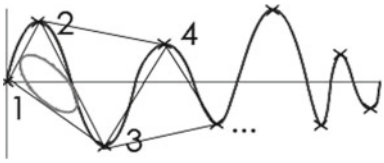


Fig. 4 Triangle and its inner elliptical formation

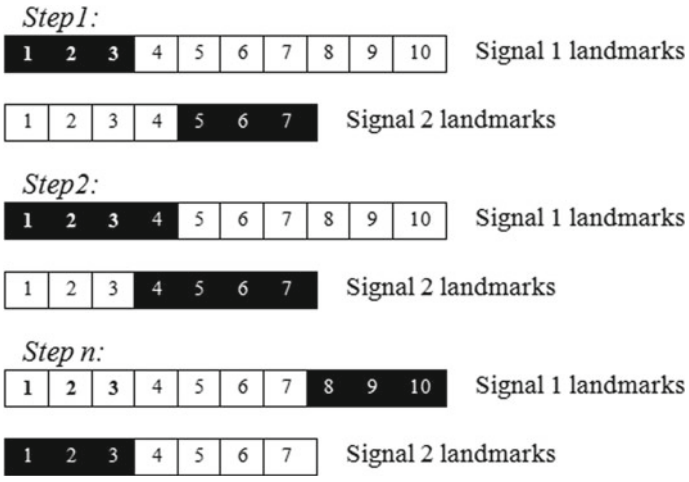


Fig. 5 Consideration of the two signals for the sphericity estimations

3.4 Sphericity Estimation

Figure 4 shows the formation of a triangle and its inner ellipse by connecting three consecutive landmarks of the signal. Figure 5 shows mainly the consideration of two signals for the sphericity estimations at each step.

Consideration of two signals for the sphericity estimations is the key role in the detection of artifact or pattern in brain signal analysis. Initially, at step 1, first 3 landmarks of signal 1 and last three landmarks of signal 2 are considered for the estimation. In step 2, first four landmarks of signal 1 and last four landmarks of signal 2 are considered. These considerations are continued till last three landmarks of the signal1 and first three of the signal 2 are considered.

Figure 6 shows the sphericity estimations at one particular consideration of two signals. This may give rise to the sequence of sphericity estimations as shown in the third column of Table 1.

At each iteration, three landmarks from both the signals are considered and estimated the sphericity of those two triangles.

Table 1 Sample sphericity estimations between two signals S1 and S2

SN	Landmarks considered	Sphericity
1	S1 (3, 4, 5) with S2 (1, 2, 3)	0.43
2	S1 (4, 5, 6) with S2 (2, 3, 4)	0.69
…	…	…
5	S1 (7, 8, 9) with S2 (5, 6, 7)	0.90

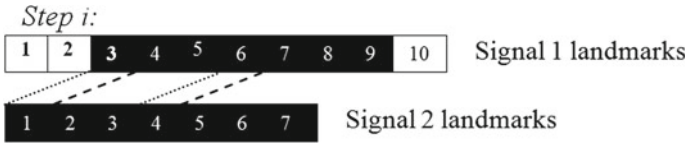


Fig. 6 Sphericity estimations between two signals

Table 2 Converted sphericity values from continuous to binary

SN	Landmarks considered	Sphericity
1	S1 (3, 4, 5) with S2 (1, 2, 3)	0
2	S1 (4, 5, 6) with S2 (2, 3, 4)	1
…	…	…
5	S1 (7, 8, 9) with S2 (5, 6, 7)	1

3.5 Postprocessing

Postprocessing converts the continuous values of sphericity to the binary that is either 0 or 1. Table 2 shows the conversion by considering threshold value of 0.60.

Equation 2 finds the total matching between two signals

$$TM2Sa1 = \frac{N1}{n} \tag{2}$$

where TM2Sa1 is the total matching between two signals at one particular consideration (gives the percentage of matching between two signals), *N1 is Number of 1's* and *n is the Total number of estimations*.

Similarly, TM2Sa1 is estimated for all other considerations of two signals. Finally, the maximum among all the consideration can be regarded as similarity index of two signals. Such similarity indices between number of signal samples are listed in the Table 3.

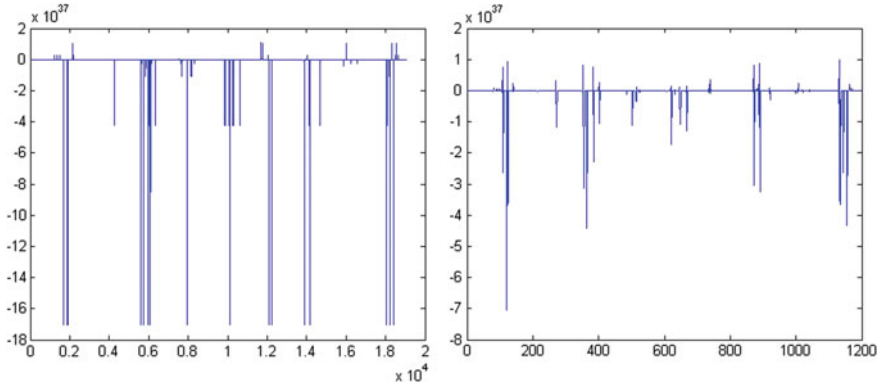


Fig. 7 Original signal captured during fourth interval of subject2 (S2_4), approximation coefficient of S2_fA_4

4 Results

The captured signal will undergo DWT and finally results fA_5 and fD_5. This experiment uses only fA_5 for similarity index estimation with other fA_5 part of other signal which is shown in Fig. 7. This paper has considered around four samples from each subject. Where the fourth subject was made to watch TV during signal capture. The similarity indices between other samples of other subject with fourth subject samples is not crossing 30%. Hence, the proposed system is considering the threshold for similarity index as 0.30. Table 3 shows all the similarity indices between signals. Some of the similarity indices are crossing 0.30, which means that somewhere while answering the questions/items every subject might have thought about the similar aspect. That may be encountered in the form of action potential during the signal capture.

The part (exactly where matching takes place) of signals whose similarity indices crossing 0.3 can be regarded as artifact or form of EEG signal pattern while answering the item which is shown in Table 3.

Each coefficient in Table 3 represents the percentage similarity index between any two signals. For example, the similarity index between S1_1 and S2_3 is 0.3778 that is 37.78% portion of both the signals at some place are similar. This can be considered as both the subjects while signal capturing has thought about solving the given item or the other parts of the signal is not matching, may be due to the other actions and thoughts. Hence, the signals between which the similarity index is crossing 30% can be regarded as EEG artifact for the selected item.

Table 3 Similarity indices between two signals of all the samples of all the selected subjects

	S1-1	S1-2	S1-3	S1-4	S2-1	S2-2	S2-3	S2-4	S3-1	S3-2	S3-3	S3-4	S4-1	S4-2	S4-3	S4-4
S1-1	1															
S1-2	0.2047	1														
S1-3	0.2484	0.252	1													
S1-4	0.2357	0.2283	0.1908	1												
S2-1	0.3488	0.3488	0.3256	0.3488	1											
S2-2	0.4634	0.3171	0.3659	0.3171	0.5122	1										
S2-2	0.4634	0.3171	0.3659	0.3171	0.5122	1										
S2-3	0.3778	0.3111	0.3333	0.3111	0.5581	0.7805	1									
S2-4	0.3585	0.3019	0.3208	0.3396	0.2385	0.561	0.5556	1								
S3-1	0.2739	0.2362	0.2254	0.2341	0.2791	0.3659	0.3111	0.2642	1							
S3-2	0.1975	0.2205	0.2081	0.2098	0.2558	0.2439	0.2444	0.2075	0.1972	1						
S3-3	0.3439	0.2913	0.2023	0.2811	0.3488	0.3902	0.3556	0.3774	0.2486	0.2216	1					
S3-4	0.2611	0.2756	0.2312	0.2537	0.4419	0.4146	0.3556	0.3962	0.2385	0.2146	0.3189	1				
S4-1	0.2675	0.2441	0.2717	0.239	0.1256	0.1171	0.1333	0.1208	0.2569	0.2192	0.1297	0.2398	1			
S4-2	0.2548	0.2047	0.2312	0.2244	0.3023	0.2927	0.2444	0.2642	0.2477	0.2466	0.2541	0.2197	0.2489	1		
S4-3	0.2293	0.2205	0.237	0.2098	0.2791	0.2927	0.2889	0.2642	0.211	0.2055	0.2216	0.2242	0.2217	0.2203	1	
S4-4	0.1312	0.2283	0.2659	0.278	0.1256	0.1902	0.1333	0.2774	0.1936	0.242	0.1459	0.2691	0.2534	0.2717	0.2379	1

5 Conclusion

This paper has tried to show that common thinking evaluations of two different or same subjects whenever a common item of any assessment is provided. It is found that around 30–40% similarities between signals of two different subjects. Hence, the portion of these signals can be regarded as the artifact of EEG signal for a particular item of any assessment tool.

6 Declaration

Authors have obtained all ethical approvals from appropriate ethical committee and approval from the subjects involved in this study.

References

1. Ciora, R., et al.: Intelligent assessment tool. In: Signals and Systems Conference, 2008 (ISSC 2008). IET Irish, IET (2008)
2. McCusker, K.A., et al.: Intelligent assessment and content personalisation in adaptive educational systems. In: 2013 International Conference on Information Technology Based Higher Education and Training (ITHET). IEEE (2013)
3. Bar-On, R., Handley, R., Fund, S.: The impact of emotional and social intelligence on performance. In: Druskat, V., Sala, F., Mount, G. (eds.), Linking Emotional Intelligence and Performance at Work: Current Research Evidence. Lawrence Erlbaum, Mahwah, NJ (2005)
4. Bar-On, R., Tranel, D., Denburg, N.L., Bechara, A.: Exploring the neurological substrate of emotional and social intelligence. Brain **126**, 1790–1800 (2003)
5. Goleman, D.: Working with Emotional Intelligence. Bantam Books, New York (1998)
6. Mayer, J.D., Salovey, P., Caruso, D.R.: Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT). Multi-Health Systems Inc, Toronto, Canada (2002)
7. Thurstone, L.L.: The mental age concept. Psychol. Rev. **33**, 268–278 (1926)
8. Kaufman, A.S.: IQ Testing 101. Springer Publishing, New York (2009). ISBN 978-0-8261-0629-2
9. Mouneshachari, S., Pande, M.B.S., Rao, T.S.S.: EQ and IQ based classification of intelligent index (S-quotient) using K-means. In: 2016 IEEE 6th International Conference on Advanced Computing (IACC), pp. 101–105, Bhimavaram, India (2016). https://doi.org/10.1109/IACC.2016.28
10. Millet, D.: Berger's invention has been described "as one of the most surprising, remarkable, and momentous developments in the history of clinical neurology". In: The Origins of EEG International Society for the History of the Neurosciences (ISHN) (2002)
11. Ma, W., Tran, D.; Le, T., Lin, H., Zhou, S.-M.: Using EEG artifacts for BCI applications. In: 2014 International Joint Conference on Neural Networks (IJCNN), pp. 3628–3635, 6–11 July 2014. https://doi.org/10.1109/IJCNN.2014.6889496
12. Strang, G., Nguen, T.: Wavelets and Filter Banks, pp. 174–220, 365–382. Wellesley-Cambridge Press, MA (1997)
13. Chan, A.K., Goswami, J.C.: Fundamentals of Wavelets, Wiley-India Edition. Wiley, New Delhi (1999)
14. Ansari, N., Delp, E.J.: Partial Shape Recognition: A Landmark-Based Approach. IEEE (1990)

15. Gans, D.: Tranformations and Geometries. Appleton-Century-Crofts, New York (1969)
16. Alamgir, M., et al.: Signal detection for cognitive radio using multiple antennas. In: 2008 IEEE International Symposium on Wireless Communication Systems, ISWCS'08. IEEE (2008)
17. Azemi, A., Stoock, C.: Utilizing MATLAB in undergraduate electric circuits courses. In: Frontiers in Education Conference Proceedings, vol. 2, pp. 592–602 (1996)
18. Mathworks Inc.: MATLAB users guide (2012). http://www.mathworks.com
19. NeuroSky, I.A.: ThinkGear socket protocol, Technical Report (2010). http://www.neurosky.com/

Design and Comparison of Electromagnetically Coupled Patch Antenna Arrays at 30 GHz

Sujata D. Mendgudle, Shreya A. Chakraborty, Jinisha Y. Bhanushali, Manmohansingh Bhatia and Sachin B. Umbarkar

Abstract This paper gives the brief comparison of four, eight, and sixteen elements electromagnetic-coupled patch array antenna at the center frequency of 30 GHz. Directivity, gain, radiation efficiency, S-parameter, 3D far-field, polar plots, and beam width are compared to get optimum performance at 30 GHz. This novel structure makes the use of electro-coupling feeding technique. The Gaussian excitation pulse is used to understand the performance of the antenna. The models have been simulated using Finite Integration Technique and add to open boundary condition at −30 dB accuracy. It is also possible to measure radiated field using E far-field and E near-field sensor probes at certain distances.

Keywords Micro-strip antenna (MSA) · Micro-strip patch array antenna (MPA) Electromagnetic-coupling · FIT

1 Introduction

Wireless communication plays a significant role in the technological advancements taking place today. There is a growing need for miniaturized devices as the communication equipments like mobile phones are constantly shrinking in size. This has led to a rising interest in smaller and low profile antennas such as micro-strip antennas (MSAs) for wireless transmission. Due to their small size and low cost, they have been of particular importance and subject of research for the last few years. Many studies have been carried out in this respect for determining the

S. D. Mendgudle (✉) · S. A. Chakraborty · J. Y. Bhanushali · M. Bhatia · S. B. Umbarkar
Department of Electronics Engineering, Ramrao Adik Institute of Technology (Affiliated to Mumbai University), Nerul, Navi Mumbai, India
e-mail: mendgudle123@gmail.com

S. B. Umbarkar
e-mail: sachin.umbarkar@rait.ac.in

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_61

efficiency of MSAs for wireless applications like WLAN, etc. which operate in the frequency range of around 2–3 GHz. However, recently the focus has been mainly upon higher frequencies in the millimeter range (EHF), so as to take the pressure off the lower frequencies. With the extension of wireless cellular services with 4G and 5G technologies, it has become necessary to expand the spectrum of wireless communications. Application of MSAs at such high frequencies hence becomes paramount. As pointed out in [1, 2], the propagation characteristics of mm-waves [30 GHz and above] make them a good candidate for future wireless communication systems.

While conducting research, our main interest was to compare the gain, directivity and scattering parameters of 4-element, 8-element, and 16-element antennas in order to determine which antenna performs best at 30 GHz so that such antennas can be used in future as part of MIMO systems.

1.1 Literature Survey

As narrow BW of MSAs [typically 1–5%] limits their application, increasing the BW of MSAs has drawn attention as an active research area in this field with a BW up to 70% already in place [3, 4].

The choice of different substrates for making patches of appropriate dimensions optimized to yield resonance within close by intervals to yield broad BW is discussed in Ref. [5]. Another efficient method used for the enhancement of patch antenna bandwidth at 2.4 GHz is the loading of micro-strip patch antenna with a trapezoidal slot [6]. In addition, proximity coupled antennas have also been designed to suppress second and third harmonics to reduce spurious radiation [7].

1.2 Feeding Concept

The method for indirectly exciting a patch employs electro-coupling is shown in Fig. 1. It is a combination of proximity coupling and coaxial feeding technique. A two substrate model is used to incorporate two layers of patches to be coupled. The lower substrate serves as a platform for the feed patches or the patches which are to be given the actual excitation via a coaxial feed line. The second or the upper substrate layer rests on this first layer providing the base for the main radiating patches of the antenna. The coupling phenomenon takes place between the upper and the lower patches and the signal is radiated by the upper patches of the antenna. An advantage gained by this feed configuration is the elimination of spurious feed-network radiation. Choice of two different dielectric media, one for the patch and the other for the feed line optimizes the performances of both patches. An increase in the BW due to the increase in the overall substrate thickness of the MSA is obtained like this [5, 8]. By using a proximity coupled patch antenna, larger

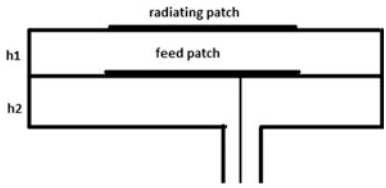


Fig. 1 Feeding structure [5, 6]

bandwidths can be realized in comparison to direct contact fed patch antennas without degrading the front-to-back ratio of the antenna [9].

1.3 Design of Electromagnetic-Coupled Arrays

Design details are as:

- Patches—PEC, Ground—PEC
- Lower substrate—Arlon AR 450 (loss free)
- Upper substrate—Arlon AD 320 (loss free).

The substrate material was chosen to be Arlon due to its excellent dielectric properties which make it suitable to be used as substrates. Also, the dielectric constants of the materials were taken into consideration while choosing the upper and lower substrates. The spacing has been taken as λ/2, i.e., 5 mm to avoid grating lobes [10]. The array was designed to have the input impedance of 50 ohm. The design parameters for the arrays are given below in Table 1.

1.4 Structural Visualization

The length of the ground plane (L) is 20.2432 mm and its width (w) is 21.9006 mm (Figs. 2, 3 and 4).

The ground plane has a length (L) of 20.2432 mm and width (w) of 38.8012 mm.

Here the length of the ground plane (L) is 35.4864 mm and width (w) is 38.8012 mm.

2 Antenna Performance and Simulation Results

Using Finite Integration Technique [FIT] as in Ref. [11], S-parameters, far-field patterns, and the antenna gain are evaluated at 30 GHz and the results thus obtained are shown below.

Table 1 Optimized design parameters

Parameters	Value
Fr	30 GHz
ε_{r1}	3.2
ε_{r2}	4.5
h1	0.335 mm
h2	0.2828 mm
Hp	0.05 mm
Wp	3.4503 mm
Lp	2.6216 mm
Wf	3.013 mm
Lf	2.258 mm

where
fr = radiating frequency = 30 GHz
ε_{r1} = Dielectric constant of Arlon AD 320 (upper substrate) (loss free) = 3.2
ε_{r2} = Dielectric constant of Arlon AR 450 (lower substrate) (loss free) = 4.5
$\lambda_{air} = \lambda_0 =$ C/fr = 0.01 m = 10 mm
λ = 5.58630 at ε_r = 3.2
λ = 4.71078 at ε_r = 4.5
h1 = Thickness of upper substrate = h1 = 0.06 $\lambda_{air}/\sqrt{\varepsilon_{r1}}$ = 0.335 mm
h2 = Thickness of lower substrate = h2 = 0.06 $\lambda_{air}/\sqrt{\varepsilon_{r2}}$ = 0.282 mm
hp = Thickness of patch (all)
Wp = width of radiating patches = Wp = C/2fr * $\sqrt{[2/(\varepsilon_{r1}+1)]}$ = 3.450 mm
Lp = length of radiating patches = Lp = [V0/(2fr * $\sqrt{\varepsilon}$eff)]-2∇L = 2.6216 mm
Wf = width of feed patches
Lf = length of feed patches

2.1 S-Parameters

2.1.1 4-Element

It is observed that the S11 parameter has its lowest peak around −54 dB at 27.5 GHz., it shows that matching at the port is quite good with less return loss. It is observed S21 which shows isolation between port 2 and 1 is approx. −40 dB at 30.5 GHz indicating that the coupling between two ports is less. The gross magnified plot of S-parameters is shown in Fig. 5a

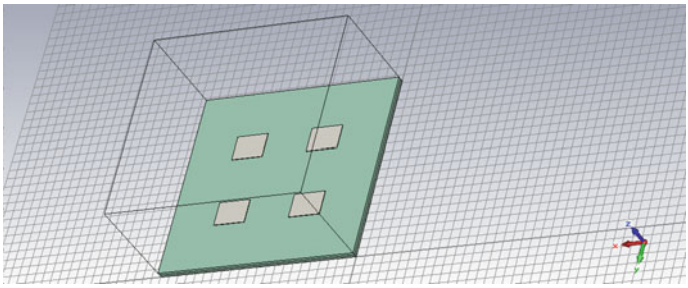


Fig. 2 4-element antenna (*4 radiating patches*)

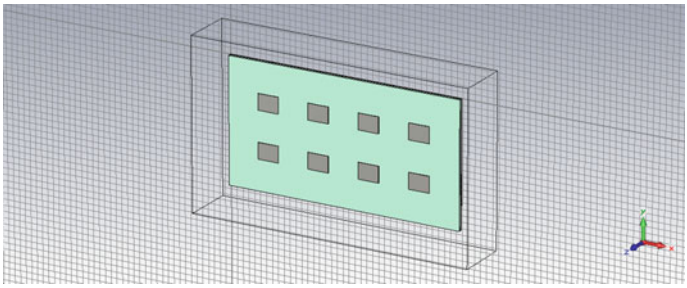


Fig. 3 8-element antenna *(8 radiating patches)*

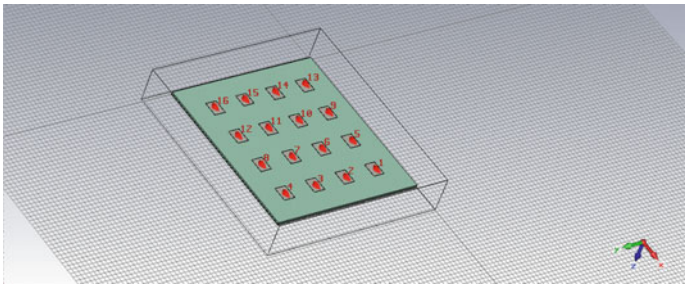


Fig. 4 16-element antenna *(16 radiating patches)*

2.1.2 8-Element

For an 8-element array, the lowest peak shows approximately −130 dB at 22.5 GHz. Fig. Fig. 5b shows the plot of gross S-parameters for different modes. It is observed that S21 is −130 dB at 26 GHz, indicating there is less coupling between port 2 to 1 showing good isolation.

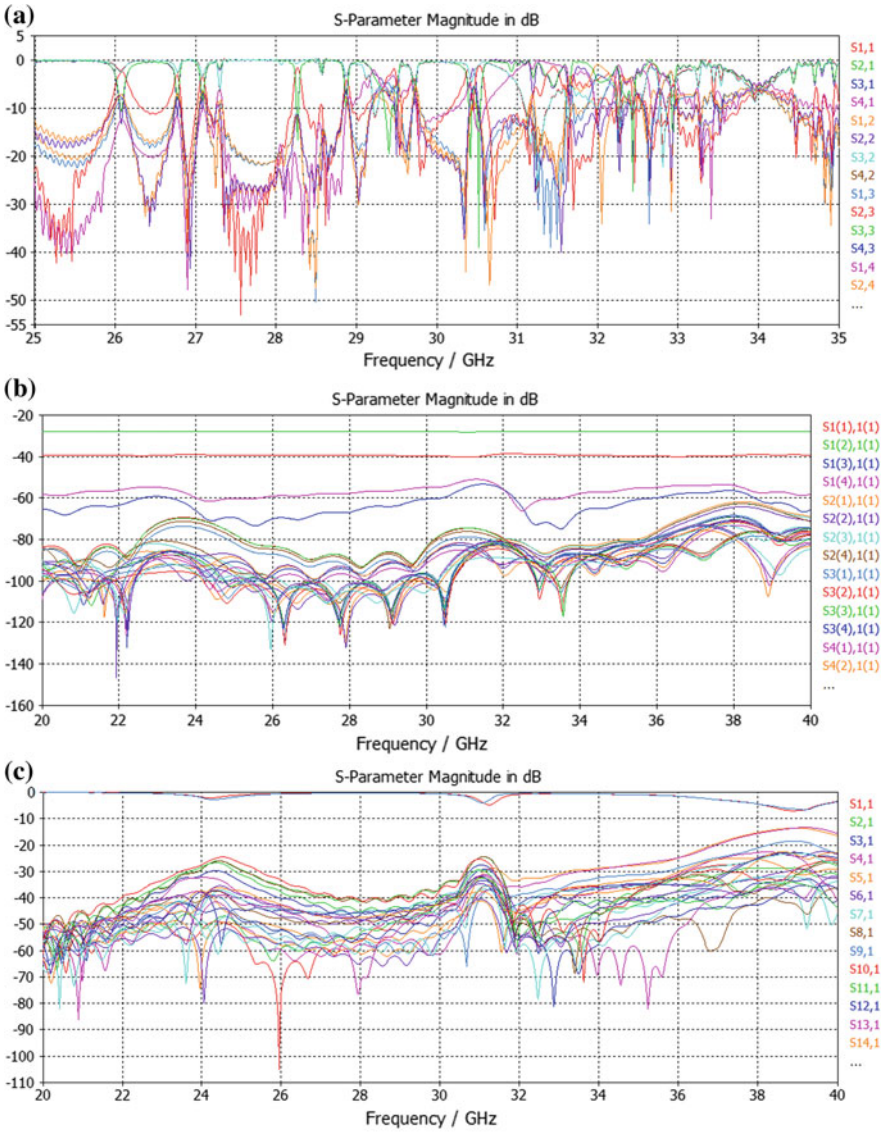


Fig. 5 **a** S-parameters for 4-element micro-strip antenna, **b** S-parameters for 8-element micro-strip antenna, **c** S-parameters for 16-element micro-strip antenna

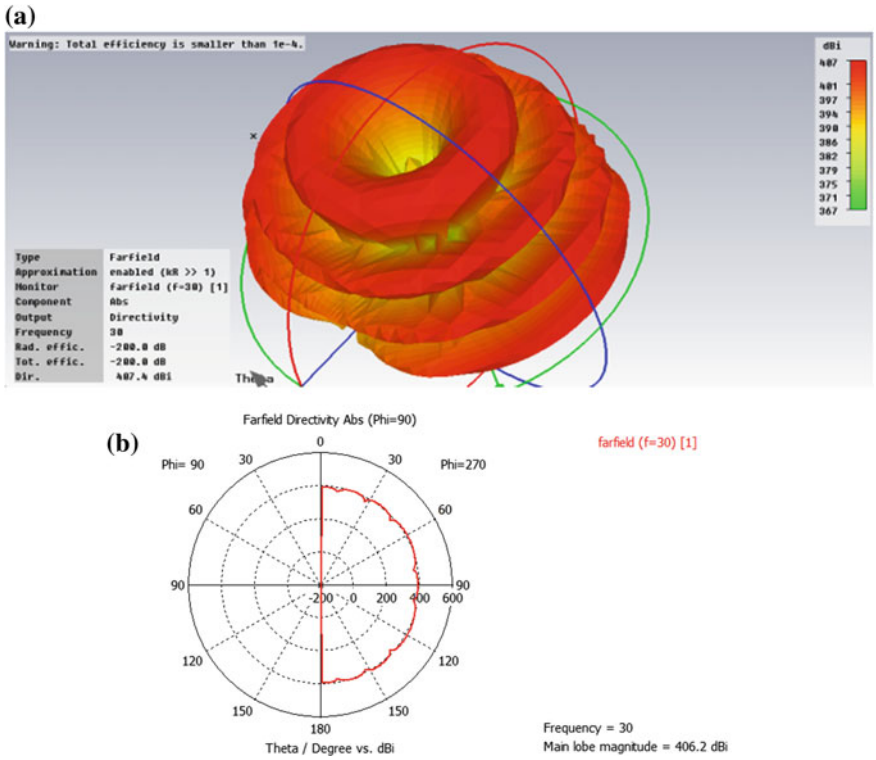


Fig. 6 **a** 3-D far field plots **b** polar plot for 4-element micro-strip antenna

2.1.3 16-Element

The lowest value of the S11 parameter is observed at 26 GHz with the value of approx. −105 dB that means port matching is quite good. The plot is shown in the Fig. 5c

2.2 Farfield Plots

2.2.1 4-Element

Figure 6a, b shows 3D Far field Plot and polar plot, respectively. We can see that at 30 GHz polar plot magnitude is 406.2 dBi and directivity 407.4 dBi with no side lobes.

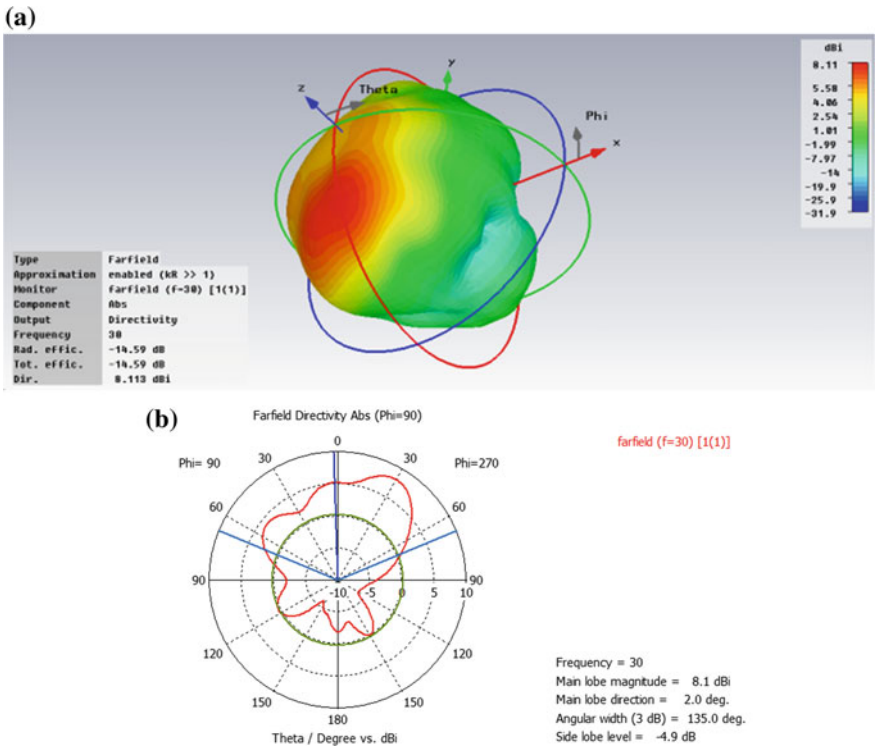


Fig. 7 **a** 3-D far field plots **b** polar plot for 8-element micro-strip antenna

2.2.2 8-Element

Figure 7a, b shows 3D Far field Plot and polar plot, respectively. Figure 7a shows that, at 30 GHz radiated efficiency and total efficiency are same equal to −14.59 dB with directivity 8.113 dBi whereas in polar plot main lobe magnitude is 8.1 dBi and side lobe level is −4.9 dB which is very small compared to the main lobe.

2.2.3 16-Element

Figure 8 a, b shows 3D Far field Plot and polar plot respectively. Figure 8 we can see that, at 30 GHz radiated efficiency −0.054 dB is more than total efficiency −8.88 dB with directivity 6.8 dBi whereas in polar plot main lobe magnitude is 6.3 dBi and side lobe level is –2.7 dB which is very small compared to the main lobe.

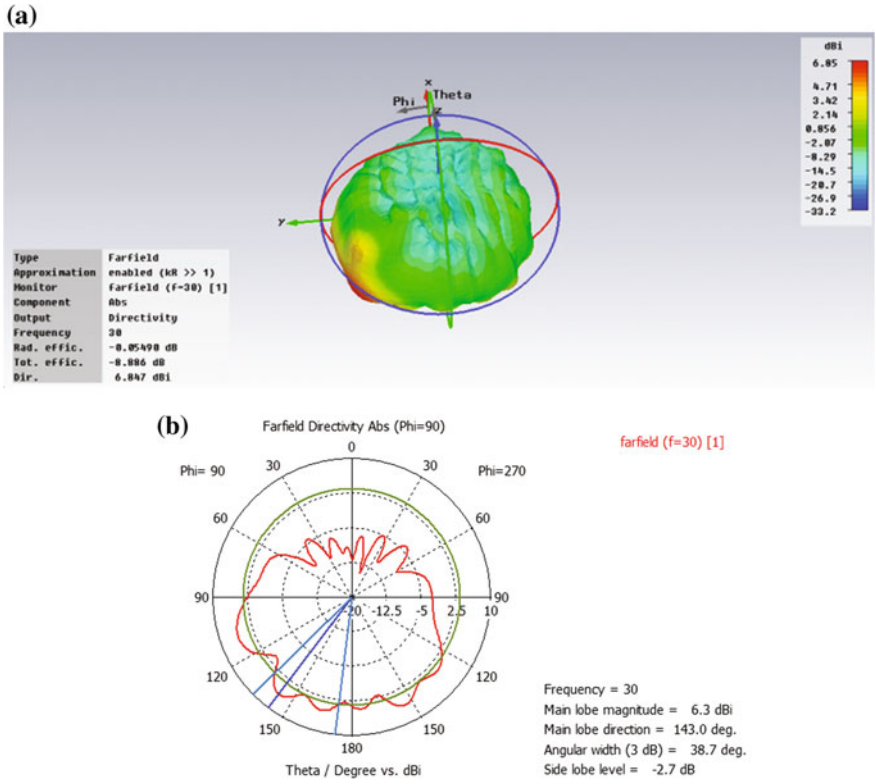


Fig. 8 **a** 3-D far field plots **b** polar plot for 16-element micro-strip antenna

2.3 *Summary of Antenna Results*

See the Table 2.

Table 2 Overview of Results

Parameter	4-element	8-element	16-element
Directivity (dBi)	407.4	8.113	6.847
Gain (dBi)	406.2	8.1	6.3
Value of S11-parameter at 30 GHz (dB)	Approx. −25	Approx. −100	Approx. −40
Lowest peak of S11 parameter (dB)	−54 dB at 27.5 GHz	−130 dB at 22.5 GHz	−105 dB at 26 GHz
Side lobe level	0	−4.9	−2.7
Radiation efficiency (dB)	−200	−14.59	−8.886
Total efficiency (dB)	−200	−14.59	−0.0549

3 Conclusion

The 4, 8, and 16-element coupled rectangular patch antennae have been designed for future applications in MIMO systems as in 5G technology. The electro-coupled feeding technique with coaxial line given to the feed patches is used to enhance the antenna's bandwidth. Antenna parameters like directivity, gain, and radiation efficiency have been observed and compared. The directivity of the 4-element patch antenna is the best among the three with the value of 407.4 dBi. Further comparing their gain, we observe the highest value at 406.2 dBi for the 4-element patch antenna. It is, therefore, concluded that among the three models designed, the 4-element design gives maximum performance at 30 GHz.

Acknowledgements The authors sincerely thank Department of Electronics, Ramrao Adik Institute of Technology, Nerul, Navi Mumbai for the constant guidance, support, and encouragement throughout. We are also grateful to Electron Beam Center (BARC), Kharghar for providing us the required platform to conduct our research.

References

1. Gunnarsson, S.E, Wadefalk, N., Svedin, J., Cherednichenko, S., Angelov, I., Zirath, H., Kallfass, I., Leuther, A.: A 220 GHz single-chip receiver MMIC with integrated antenna. IEEE Microwave Wirel. Compon. Lett. **18**, 284–286 (2008)
2. LiQuan, H.: Some advances in millimeter wave application systems. In: Proceedings of 1997 Asia-Pacific Microwave Conference
3. James, J., Hall, P.: Handbook of microstrip antennas. P. Peregrinus on behalf of the Institution of Electrical Engineers, London, U.K. (1989)
4. Lee, K.F, Chen, W.: Advances in microstrip and printed antennas. Wiley, New York (1997) ISBN 978-0-471-04421-5
5. Kumar, G., Ray, K.P.: Broadband microstrip antennas, pp. 16–17. Artech House, Boston (2003) ISBN 1-58053-244-6
6. Hamad, K.: Design and enhancement bandwidth rectangular patch antenna using single trapezoidal slot technique. ARPN J. Eng. Appl. Sci. **7** (2012)
7. Inclan-Sanchez, L., Vazquez-Roy, J., Rajo-Iglesias, E.: Proximity coupled microstrip patch antenna with reduced harmonic radiation. IEEE Trans. Antennas Propagat. **57**, 27–32 (2009)
8. Rathi, V., Kumar, G., Ray, K.P.: Improved coupling for aperture coupled microstrip antennas. IEEE Trans. Antennas Propagat. **44** (8), (1996)
9. Emhemmed, A., McGregor, I., Elgaid, K.: 200 GHz broadband proximity coupled patch antenna. In: 2009 IEEE International Conference on Ultra-Wideband (2009)
10. Balanis, C.: Antenna Theory—Analysis and Design (3rd edn.), p. 310. Wiley
11. Clemens, M., Weil, T.: Discrete electromagnetism with the finite integration technique. Prog. Electromagn. Res. **32**, 65–87 (2001)

Part VI
Internet, Web Technology, IoT, and Social Networks & Applications

Natural Language Query to Formal Syntax for Querying Semantic Web Documents

D. Suryanarayana, S. Mahaboob Hussain, Prathyusha Kanakam and Sumit Gupta

Abstract The impact of the search engines has grown along with World Wide Web and it becomes the central role for the knowledge engineering. Semantic Web offers a pathway to process the knowledge and the enormous data from the documents on the Web which comprise semi-structured data. This paper proposes some strategies for execution of the queries on the Web of data to repossess the knowledge and information by the Web search engines. Natural language processing will be applied to convert the natural language queries to formal syntax to retrieve the semantics and pragmatics of the data from the existed ontologies on the Web to provide accurate information to users.

Keywords Semantic web · NLP · Ontologies · SPARQL · RDF Pragmatics

1 Introduction

Nowadays the important task is to provide precise information to the users' from the Web of various data formats. Web crawlers are foreseen as an important part to gather and collect the data from the network of documents. Rather than the traditional Web architecture, semantic Web plays a major role to provide accurate information for a natural language query by using special query languages on

D. Suryanarayana (✉) · S. M. Hussain · P. Kanakam · S. Gupta
Department of Computer Science & Engineering,
Vishnu Institute of Technology, Vishnupur, Bhimavaram, Andhra Pradesh, India
e-mail: suryanarayanadasika@gmail.com

S. M. Hussain
e-mail: mahaboobhussain.smh@gmail.com

P. Kanakam
e-mail: prathyusha.kanakam@gmail.com

S. Gupta
e-mail: sumit108@hotmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_62

the ontologies. Integration of the data has been another imperative issue for the Web and Internet-based data frameworks–it is difficult to consolidate new data with whatever the related data that currently available, and to make them both accessible for the queries. The integration of the ontologies will make possible to build a promising application for searching and querying to understand the intent of the user queries. The advancement of the current Web is the semantic Web where the information is well described in semantics [1].

Semantic Web enables the search engine to interpret Web contents in the same way as a human being while it is fast and accurate. Semantic search attempts to extend and improve traditional search results. The main aim of the semantic search is to understand cognition of user query. The normal search engine gives the results based on the keywords but a Semantic search will take into account the context and meaning of search terms. It understands the cognition of the searcher when typing in that search query. Ontology is required to develop the semantic search engine and used to describe the domain knowledge and ease to get accurate or more precise results to the user query [2].

2 Preliminary

2.1 Ontology

Ontology is expressed with a well-defined set of classes and properties descriptions and it is an analytical study to acknowledge the things, acceptable, reality, or phenomenon and the primitive types of the things and their associations. Resource description framework (RDF) has turned into defacto standard for the semantic data and ontologies. Ontology regularly deals with the queries regarding the being prevail or to be mentioned, and the process of organizing the things, its associated pecking order and segregated appropriately to the likeliness and unlikelihood. The field of Semantic Web creates ontologies to limit complexity and to organize information. It is a conceptual representation of domains, appropriate approach and in which concept they belong. Ontologies can be represented via classes, relations, and instances. There are different levels of formalizations for describing ontologies from informal range to formal range. As per the level of the majority, there exist distinct categories of ontologies.

- Peak level or upper or foundation level ontology represents general concepts and anything in the world can be classified.
- Domain ontology focus on a specific domain which is defined in an upper ontology is described more specifically.
- In task ontology, fundamental concepts like task and general activity are described.
- Application level ontology focuses on specific task and domain.

2.2 Resource Description Framework and Schema

Knowledge can be represented more expressively using resource description schemas (RDFS) which are an RDF vocabulary description language used to illustrate vocabularies for RDF, means the models that are dealing [3]. See Figs. 1 and 2.

Definition 1 (*RDF terms*) Let us consider the set of all *IRIs* (refers to the resource description framework, Web ontology language, and XML schema) as I_r for *literals set* L_r, and set of all *nodes* N_b, then the *RDF terms set* defined as $T = I_r \cup L_r \cup N_b$.

RDFS consent to the description of Classes via *rdfs:Class* and creation of a real instance of a class in RDF via *rdf:type,* for example*:Planetrdf:typerdfs:Class, Earthrdf:type:Planet.*

Here planet is an *rdf:type* class and earth is the resource which has type planet. Therefore, the earth is an instance or member of a class. Besides classes, it defines the properties of classes. Properties connect with classes either with literals. It can define a property via *rdf:Property*. On property, one can define restrictions on domain and range according to type via *rdfs:domain* and *rdfs:range*.

In *rdf schema* everything discusses resources and it defines hierarchical relationships like sub-classes, super-classes, and sub-properties and super properties. Ontologies are developed as a further semantic Web standard as a reason that the RDF will not accomplish more complicated constraints concerning properties of the classes and the resources. The below is a semantic web document of a person details written in RDF/XML.

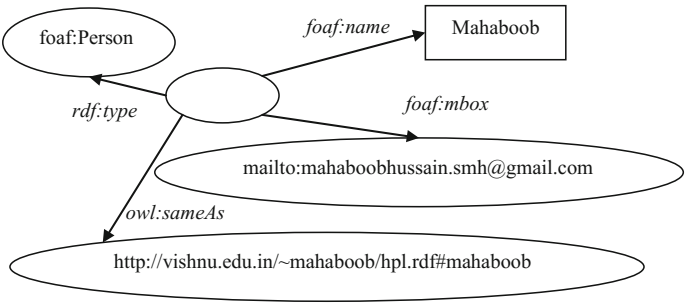


Fig. 1 The RDF representation of the example of *foaf:Person* from the above semantic Web document

```
1: <?xml version="1.0" encoding="utf-8"?>
2: <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
syntax-ns#"
3:            xmlns:owl="http://www.w3.org/2002/07/owl#"
4:            xmlns:foaf="http://xmlns.com/foaf/0.1/">
5:      <foaf:Person>
6:      <foaf:name>Mahaboob</foaf:name>
7:      <foaf:mbox
rdf:resource="mailto:mahaboobhussain.smh@gmail.com"/>
8:      <owl:sameAs
rdf:resource="http://http://vishnu.edu.in/~mahaboob/foaf.
rdf#mahaboob "/>
9:      </foaf:Person>
10: </rdf:RDF>
```

The above document represented as RDF graph as in Fig. 1 with all details of the person mentioned in the document.

2.3 Protocol and RDF Query Language (SPARQL)

To manipulate and to retrieve the data which is stored in Resource Description Framework format, an RDF query language SPARQL is required. It indicates the rules, patterns, and the significance of the SPARQL query language for RDF. The result sets and RDF graphs are the results of SPARQL queries that execute on RDF [4]. The query form of a SPARQL allows a triplet {*subject, predicate, object*} pattern, union, intersection strings, and various elective patterns. The modification process is not completely reliant on the means of the query and the result sequence modifiers of SPARQL. Thus, the whole documents in the database are in the form of triples {*subject, predicate, object*}. The SPARQL endpoint is an RDF triple database on the server usually which is available on Web and top of Web transfer protocol there is a SPARQL protocol layer means via HTTP SPARQL query is transferred to server and server gives its results to a client. It is like SQL but works on RDF graphs not on tables [5].

Definition 2 (*RDF graphs*) The *graph pattern* P_g is the member of a *triplet* pattern set *(subject, object, predicate)* defined as $(T \cup V) \times (I_r \cup V)?(T \cup V)$ or *(subject, object, predicate)* $\in (I_r \cup N_b) \times I_r \times T$.

The graph pattern is RDF triple that contains some patterns of RDF variables. Patterns can be combined to get different patterns of more complex results.

Definition 3 (*Solution multiset* (P_g)) A limited function defined from the *blank nodes* in the *RDF* terms as $\sigma \sim N_b \to T$. Let illustrate the various *mappings* of *RDF* as $\sigma_1, \sigma_2, \ldots \sigma_n$ then, $\{m(\sigma_i = V(P_g).\ \mu(\sigma_i(P_g))$ is a *subgraph* of a graph *G,* $\forall$ $1 \le i \le n$ where η is the utmost number in $(\mu, \eta)|som(\mu) = V(P_g)$ and V are the countable infinite set*}*.

3 From Natural Language Query to Formal Query

Machine need to understand the intent of the users' query posed to the search engine to retrieve the accurate results. Semantic Web can handle to retrieve the semantics from the queries with SPARQL syntax from the ontologies constructed with triplet pattern RDF. Answering the semantic queries is a high priority task for the semantic search engines with the efficient Web semantic crawlers to retrieve the semantic Web documents [SWDs] from the database which already existed. Design and construction of ontology are quite easy with the help of protégé tool [6]. Querying on the ontology generated with the RDF is a critical task which employs syntax-based formal query language. Here in this paper, SPARQL query language is used to retrieve the information from these ontologies.

It is easy to retrieve the information from the ontology-based database by using a SPARQL query language, but the application which takes a query in the form of natural language can be converted into the formal query (SPARQL). This paper presents the process of converting the natural language into the formal query which can post on the ontology directly to retrieve the information. For example consider this query (Q): *"what are the publications of Mahaboob while working in the college Vishnu Institute of Technology?"* querying the Vishnu academic ontology. Initially, it can be converted as a conjunctive formal logic expression as:

$$?x(?x\,is, publications) \backslash (Mahaboob, workingIn, ?x) \backslash (?x, college, Vishnu\ Institute\ of\ Technology)$$

where, class *publications,* instance *Mahaboob* and *Vishnu Institute of Technology* are the nodes which consist of ontological limitations while *workingIn* and *college* are the required connecting arcs pattern.?*p* is a variable that passes constraints on the query. Thus interpreting the natural language to a formal query will be made easy via these formal logics. Assume that the above query in SPARQL format applied to the knowledge base as shown below will result in the related information.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-
ns#>
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
@prefix ex: <http://example.org/Publications#>
ex:Mahaboob_Hussain:typeex:Writer
ex:Mahaboob_Hussains:label""Mahaboob Hussain""@sv
ex:Prathyusha_Kanakam:typeex:Writer
ex:Prathyusha_Kanakam:label""Prathyusha Kanakam""@sv
ex:Stepping_Towards_Semantic_Search_Enginerdfs:label""Ste
pping Towards Semantic Search Engine""@en
ex:Stepping_Towards_Semantic_Search_Engineex:authorex:
Mahaboob_Hussain
ex:Stepping_Towards_Semantic_Search_Engineex:numberOfPage
s""09""^^<http://www.w3.org/2001/XMLSchema#int>
ex:Stepping_Towards_Semantic_Search_Enginerdfs:label""Ste
pping Towards Sematic Search Engine""@en
ex:Stepping_Towards_Semantic_Search_Engineex:authorex:
Mahaboob_Hussain
ex:Stepping_Towards_Semantic
_Search_Engineex:numberOfPages""06""^^<http://www.w3.org/
2001/XMLSchema#int>
ex:E_Nose:label""E Nose""@de
ex:E_Nose:authorex:Prathyusha_Kanakam
```

The above query (ϱ) represented as below:

```
SELECT ?publicationName WHERE {
?publication ex:authorex:Mahaboob_Hussain
?publication rdfs:label ?publicationName
}
```

Then, the publication of the author for the above query is the result as below:

```
SELECT ?publicationName WHERE {
?publication ex:authorex:Mahaboob_Hussain
?publication rdfs:label ?publicationName
}
```

Therefore, by using this query format it is easy to retrieve information from the ontology database in a semantic manner.

4 Conclusion

Converting the natural language sentence into a perfect formal query format to retrieve the information on a particular database is a critical task. Since the search engine needs to predict the cognition of the human natural language sentences and alter the query into machine understandable syntax to search the semantic web accordingly. This paper deals with a sample query in the traditional natural language and its conversion into the formal syntax query and then into the SPARQL query format. Thus, any natural language sentences posted to the semantic search engine application will be easily converted into the formal syntax. So, the performance of information retrieval increases on the semantic Web documents for accurate and consistent information to the users.

Acknowledgements This work has been funded and supported by the Department of Science and Technology (DST), Govt. of India under the Grants No. SRC/CSI/153/2011.

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Sci. Am. **284**, 34–43 (2001)
2. Dai, W., You, Y., Wang, W., Sun, Y., Li, T.: Search engine system based on ontology of technological resources. JSW **6** (2011)
3. Marzano, G.: Using resource description framework (RDF) for description and modeling place identity. Procedia Comput. Sci. **77**, 135–140 (2015)
4. Kim, K., Moon, B., Kim, H.: R3F: RDF triple filtering method for efficient SPARQL query processing. World Wide Web **18**, 317–357 (2013)
5. Jeon, M., Hong, J., Park, Y.: SPARQL query processing system over scalable triple data using SparkSQL framework. J. KIISE **43**, 450–459 (2016)
6. Segaran, T., Evans, C., Taylor, J.: Programming the Semantic Web. O'Reilly, Beijing (2009)

Bat Inspired Sentiment Analysis of Twitter Data

Himja Khurana and Sanjib Kumar Sahu

Abstract The paper aims to implement a supervised learning approach to perform sentiment analysis. The data has been classified into following: positive, neutral and negative. The main objective is to discover the efficiency of an adapted version of the Bat Algorithm. Input textual data has been extracted from a live stream of 'tweets'. We compare two models in the paper: Unigram and Unigram with POS tags through empirical evaluation of results expressed in terms of standard metrics. The results show marginal improvement over SVM classifier. This is the first application of Bat Algorithm to the analysis of sentiment taking input from 'tweets', to the best of our knowledge.

Keywords Machine learning · Meta-Heuristic · Bat algorithm (BA) Sentiment analysis (SA) · Classification · Twitter · POS tagger

1 Introduction

Sentiment analysis has been garnering interest for more than a decade now. With an upsurge in social media usage and online expression of human experiences or opinions, the subject of SA has become the focus of NLP researchers around the world. Coupled with a recently increased commercial interest in 'social media monitoring and analysis', this research area which was earlier known as opinion mining or a branch of information retrieval, is enjoying the inflow of research endeavours at an unprecedented pace. Various experiments conducted so far can be

H. Khurana (✉)
University School of Information and Communication Technology,
Guru Gobind Singh Indraprastha University, Dwarka, New Delhi, India
e-mail: Hkhurana11@gmail.com

S. K. Sahu
Department of Computer Science and Application, Utkal University,
Bhubaneshwar, Odisha, India
e-mail: sahu_sanjib@rediffmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_63

broadly classified into two types: approach based on the Lexicon and approach based on the machine learning. Most of the research in the area utilizes the former. In this paper, we attempt to implement the latter by applying an adapted version of a popular meta-heuristic algorithm. The research takes input from a widely used micro-blogging website: Twitter, which serves as a credible database for the task at hand.

1.1 Sentiment Analysis

Classification of input (words/phrases/sentences/documents) into labelled classes, for e.g. positive, neutral and negative, on the basis of the emotion they indicate is termed as sentiment classification. A related field is text classification which is considered simpler as the classification is based directly on the words constituting the sentence. On the other hand, sentiment classification is a more subtle and complex task as the polarity of the sentence can be either opposite or un-relatable to its word constituents, for e.g. 'I wish I could go to france and meet president obama haha' or 'I think I should go to sleep'.

1.2 Twitter

Twitter is a prominent online social interaction and information dissemination platform. 332 million people use the platform to express their opinions, ideas, feelings, feedback, reviews, etc. through 'tweets' in 140 characters or less. Users can also re-tweet, follow, share or like. Currently, Twitter users post around 6,000 tweets per second on a regular basis. It corresponds to over 350,000 tweets per minute and 200 billion tweets per year. With such voluminous amount of real-time data produced at a consistent pace, twitter is a goldmine for a diverse set of applications, such as opinion mining, companies seeking feedback on their products [1] and analysis of sentiment towards a matter of social or political concern [2].

1.3 Machine Learning Approach

Research endeavours in sentiment analysis exploit one of the following aspects: lexicon features and machine learning techniques. While the former chooses to improve efficiency by proposing new ways of knowledge representation and inclusion of additional 'features' along with input data, latter explores the development of algorithms to improve the accuracy of the task. More generally, machine learning is the study and construction of algorithms that can learn from and make predictions on data. Earlier examples include testing the efficiency of genetic

algorithm and particle swarm optimization for different domain areas. In this experiment we utilize certain parameters of bio-inspired bat algorithm to perform sentiment analysis.

1.4 Bat Algorithm

The new generation algorithms are inspired from nature. Almost all of them have shown lucid evidence of improved efficiency and accuracy in performance. Bat algorithm proposed by Xin She Yang in 2010 is one meta-heuristic algorithm that has found much acceptance among researchers for image processing, multi-modal optimisation, clustering and classification applications. The algorithm has been inspired from movement and hunting behaviour of micro-bats. These bats use echolocation to detect their prey in the night. They emit loud signals with a high sound pulse and listen for its echo. They wait for the sound to bounce back from objects present in the surroundings. With the help of this echo, they calculate their distance from a prey and also identify whether the object is a prey or an obstacle. Their emitted sound pulses vary in properties as per type of species. We will be assigning a polarity to input text using a bat—inspired technique. The standard bat algorithm has five parameters, viz. velocity v_i, position $x_{i,}$ loudness $A_{o,}$ frequency f_{min} and wavelength λ. However, we will be using only three parameters for this experiment—frequency, loudness and position.

2 Literature Review

2.1 History of Sentiment Analysis

The earliest work in the field perhaps dates back to 2003 [3], and the term opinion mining was used in the context of product reviews extraction [4] for the first time. Though, the research work on sentiments and opinions appeared earlier [5–9]. Most of the early work focuses on genre categorization and subjectivity detection [10–12]. But these techniques were not helpful in identification and assignment of the polarity of the sentence. In the past, there have been several successful attempts to perform SA. Yet most of them were based on the lexical approach [13, 14, 15, 16, 2, 17]. IIT-B (India) has contributed substantially to the subject [18, 19, 16]. Recently, an interesting case study came to light when a team of students at the esteemed institute helped a new and budding political party gain an absolute majority in prominent state elections through analysis of public sentiment towards its manifesto on online platforms. Needless to say, there is scope for improvement as there are no means yet to identify myriad of sentiments like anger, humour, disappointment, satire, etc.

2.2 Previous Applications of Bat

Since its proposal in 2010 bat algorithm [20] has been implemented to a wide range of applications that include optimization and classification problems [21]. Initial applications of BA include continuous optimization of engineering design which showed that BA can solve nonlinear problems quite efficiently [22]. A study of using BA on combinatorial problems was presented based on the combined economic load and emission dispatch problems. It concluded that the BA is not only easy to implement but also shows higher efficiency and accuracy in comparison to many existing algorithms [23]. A study of clustering problem for office workspaces showed a fuzzy variant of BA in contrast with PSO and GA and presented positive results [24]. On the other hand, experiments were also conducted for classification of micro-array data [25] and feature selection [26] that showed encouraging outcomes.

2.3 Use of Twitter as a Dataset

Over the years, the interest has gradually shifted from static textual data to streaming of real-time online data through social media platforms to reflect credibility, usability and relevance. Data from Twitter, particularly, has been utilized for several experiments like to polarize movie blogs [13], predict election winners [27, 28, 14], classify news stories [29], understand brand image in contrast to competitors, among others [30]. The use for sentiment analysis could be aimed towards a specific product or domain [31, 6, 10] or classification of sentiment in general [16, 7, 17, 32, 30]. Either way, it was observed that a major factor contributing to the accuracy of SA, was the diversity and ratio of the total population on the online platform. In addition, data sourced from twitter requires a laborious procedure of pre-processing wherein one has to remove and (or) replace acronyms, slangs, hashtags (#), emoticons, @ labels and spelling mistakes, to name a few. Hence, NLP scientists and researchers have suggested a number of ways to improve the accuracy of the existing methods[33, 34, 2, 17, 32]. One of these is using semantic knowledge and discourse analysis techniques which demonstrates a significant improvement in the accuracy [35, 15, 4, 8].

3 Proposed Framework

In the proposed framework we divide the data into two parts: training data (65%) and test data (35%). Tweets are extracted using a twitter API OAuth, after which they undergo a time-consuming pre-processing phase. Tasks in this phase include and are not limited to the following: special symbols {for e.g. @, #, ☺ :I etc.} are

Fig. 1 Sequence of phases in the proposed framework

removed, acronyms {for e.g. lol, rotl, bff etc.} are replaced by their respective full forms {for e.g. laugh out loud, roll on the floor laughing, best friends forever, etc.}, spelling mistakes {for e.g. tomoro, goin, becoz etc.} are corrected {for e.g. tomorrow, going, because}, etc. The data is then passed through a prototype of the classifier and results are recorded in terms of globally accepted standard metrics: Recall, Precision and F-Measure. Figure 1 presents a diagrammatic flow of the proposed model. In the proposed bat algorithm based classifier, we calculate the polarity of the sentence by cumulating the no. of occurrences of its word constituents in positive, negative or neutral sense in absolute terms and dividing it by the words' overall total no. of occurrences. We regard the sentiment classes as prey, P_i {i = −1, 0, 1} and each input row as a bat, B_i {i = 1, 2, …., n}. We calculate the distance of each bat from the prey and assign the prey with minimum distance to the bat.

4 Bat Behavior

Bat emits a signal with random loudness which are deflected back by any prey present in its surroundings. The distance is calculated by the bat by listening to the echo of the signal. The bat moves in the direction of the minimum distance prey. Bat continues to do so until the distance between his position and the prey is zero. Following methods have been used to calculate frequency, distance and position. In the domain at hand, we consider the input sentence a bat and sentiment classes prey.

4.1 Defining the Fitness Function

The Fitness function, in this case, is the word percentage that is not present in the training dataset. For each word, we check if the word is present in the training data. If for a given sentence more than 30% of words in the sentence are not present in the training set, we reject the sentence. We aim to minimize the fitness function. We have evaluated input data taking three thresholds of Fitness function—0.1, 0.2 and 0.3.

4.2 Calculation of Frequency

Given below is the equation we used to calculate the frequency of the signal produced by each bat in the training set. In other words, no. of occurrences of each word with respect to each of the three sentiment classes is calculated. Here f_{kj} represents the frequency of word k associated with sentiment j. O_{kj} represents total no. of occurrences of word k associated with sentiment j. Frequency calculation is performed using training data.

$$f_{kj} = \sum O_{kj}\{k = 1..m*n, j = -1, 0, 1\} \quad (1)$$

4.3 Calculation of Loudness

Loudness in our case is the summation of normalized ratios of frequency (f_{kj}) in a particular sense to total frequency (combining all occurrences) for each word in the sentence. We calculate this value A_{ij} for each bat/sentence i and with respect to each prey/class j.

$$A_{ij} = \sum l = 1 \text{ to } w(f_{kj} / \sum 1 \text{ to } m\, f_k) \quad (2)$$

{w = no. of words in sentence to be classified, i = 1.. m, j = −1, 0, 1}

4.4 Update Position of Bat

The position of the bat is updated iteratively by calculating the distance between the bat and the prey. In other words, the ratio of the sentiment class for which the word has been used most to its total number of occurrences is calculated towards each prey/class. The class with the highest ratio is the prey with minimum distance. The bat is assigned to that prey.

4.5 Algorithm

Split manually annotated pre-processed data into two −65% training set and 35% test data. The dataset is a knowledge matrix of size m * n; m rows and n columns. Each of the m rows is a bat B_i for i = 1 m. Each of the three classes is a prey P_j for j = −1, 0, 1. The fitness function F is the percentage of word absence in the training set.

```
Define fitness F_max = 0.3
For each Bat B_i i = 1, 2…m
    Calculate fitness F_i
    if (F_i < F_max)
        For each prey P_j = -1, 0, 1
            For each word k in the Bat B_i
                Calculate frequency f_kj using eqn. (1)
        Calculate loudness A_ij using eqn. (2)
        Adjust the position x_i
        Sentence Classified
    else Reject Sentence
```

5 Experiments and Results

5.1 Dataset Used

The data used in this paper has a few unique features. (1) Large size (2) Multilingual data (3) Input from users around the world (4) Manually annotated (5) Balanced sampling (6) Collected from live stream and (7) Minimized bias. The data has been sourced from Apoorv Aggarwal of Columbia University, USA. It was originally provided by a Next Gen Invent Corporation, a Business Intelligence Company and a commercial source for such data. They ensure the data is unbiased by gathering tweets in live streaming fashion without the use queries. It comprises of tweets uploaded by users from around the world in different languages. In fact, some of the tweets which were originally in foreign languages have been translated into English using Google Translate. We acquired three resources from the same entity [36]. One of them was a collection of 5,127 tweets manually annotated as positive, negative and neutral. Table 1 presents specifications of the dataset. The other two resources were an emoticon dictionary and an acronym dictionary which we explain in further text.

Table 1 Details of the input dataset

S. No.	Number of tokens	No. of occurrences
1.	Stopwords	30,371
2.	English words	23,837
3.	Punctuation marks	11,584
4.	Capitalized words	4,851
5.	Twitter tags	3,371
6.	Negations	942
7.	Other tokens	9,047
8.	Total tokens	79,152

Table 2 Sample of emoticon dictionary

Emoticon	Polarity
D: D8 D; D = DX v.v	Negative
:-P :P XP :-p :p = p :-	Positive
ZZzzz… (X_X) x_x	Neutral

Table 3 Sample of acronym dictionary

Acronym	Full form
10q	Thank You
12b	Wannabe
1dering	Wondering

In addition to this data we procured and used two more resources [36]: First, an emoticon dictionary. The emoticon dictionary contains symbols and corresponding polarity of 170 popularly used emoticons. The list has been prepared from Wikipedia. Table 2 presents a sample. A second resource is an acronym dictionary. It contains full forms of 5, 184 acronyms that are frequently used online. It has been compiled from an online portal.[1] Table 3 shows a small sample.

We have used two models to evaluate the proposed classifier: Unigram and Unigram + POS. The unigram model is popularly known as the 'Bag–of–Words' model in which each word is independent. The name suggests that the order of words in a sentence is not important. Each column in the model signifies the presence or frequency of a word in an input sentence. We may point out that we have not taken into account semantic features for the purpose of this experiment. For Unigram + POS model, words have been accentuated with their respective part-of-speech. We used Stanford POS tagger for this purpose. The tagger has global acceptance and proven accuracy in SA applications.

5.2 Results

The results have been empirically evaluated through standard metrics—Precision, Recall and F-measure with variations in Fitness function of the proposed algorithm. Figures 2, 3 and 4 present the obtained values for Precision, Recall and F-measure, respectively. Graphs in the figure present three point fitness cut off carried out for a comparative analysis. Tables 4, 5 and 6 show the results in a tabular format. To put things in perspective, we may state that our best results show marginal improvement over results on the same set of data using SVM [36].

[1] http://www.noslang.com

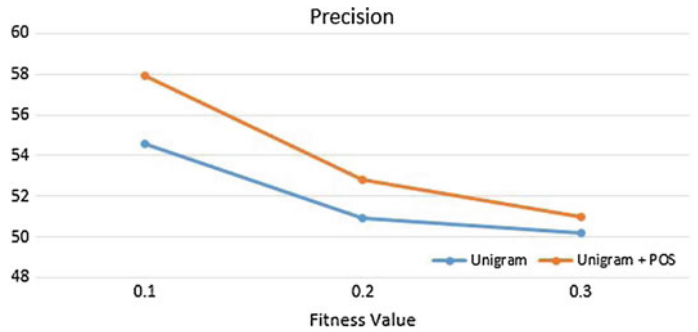


Fig. 2 Precision values for Unigram and Unigram + POS models with varying Fitness Value

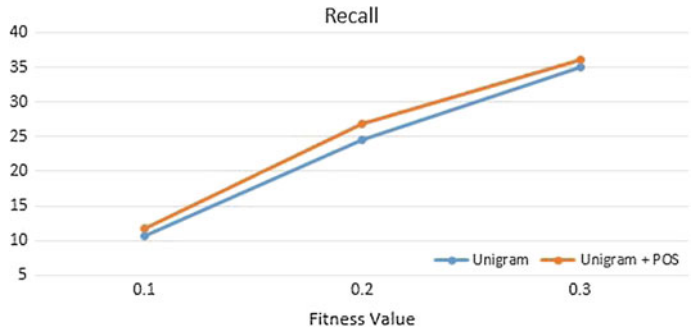


Fig. 3 Recall values for Unigram and Unigram + POS models with varying Fitness Value

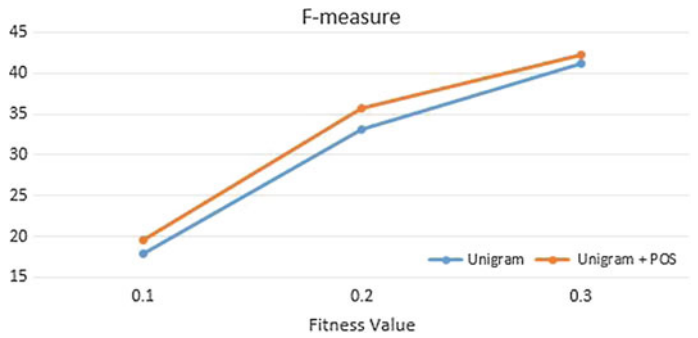


Fig. 4 F- Measure values for Unigram and Unigram + POS models with varying Fitness Value

Table 4 Precision values for Unigram and Unigram + POS models with varying Fitness Value

Fitness Value	Unigram	Unigram + POS
0.1	54.545	57.894
0.2	50.922	52.787
0.3	50.191	51.002

Table 5 Recall values for Unigram and Unigram + POS models with varying Fitness Value

Fitness Value	Unigram	Unigram + POS
0.1	10.657	11.722
0.2	24.511	26.909
0.3	34.991	36.145

Table 6 F—measure values for Unigram and Unigram + POS models with varying Fitness Value

Fitness Value	Unigram	Unigram + POS
0.1	17.830	19.497
0.2	33.0935	35.647
0.3	41.234	42.307

5.3 Discussion

In our experiment, we found that our adapted version of bat algorithm performed marginally better than SVM. On the same dataset, SVM produces an average accuracy of 56.58% for unigram model [36] while our algorithm gave 57.89%. We may mention that we have not included semantic information in our dataset. We believe the inclusion of semantic features can significantly improve the accuracy of the classification, though it is out of the scope of this experiment. Also, results may vary if loudness is decreased. During our experiment, we also found that accuracy can be greatly improved if the number of sentiment classes is reduced to only two: positive and negative, although we do not show it in this paper. We also observed that removal of Stop-Words made an only marginal difference in precision.

6 Conclusion and Future Scope

The paper illustrates the application of a Bat Algorithm inspired classifier to perform sentiment analysis of Twitter data. Empirical evaluation of the results obtained through the study show improved efficiency over SVM. Further study in the area can aim for achieving following goals:

- Testing efficiency of the proposed algorithm for sentiment analysis of data in other natural languages like Hindi, Oriya, etc.

- Testing performance of other meta-heuristic techniques like an immune algorithm.
- Testing effectiveness of the proposed algorithm when merged with other existing techniques of feature selection to develop a hybrid model.

Acknowledgements We are grateful to Apoorv Aggarwal and Next Gen Invent Corporation to source us the data used for the experiment.

References

1. Morinaga, S., Yamanishi, K., Tateishi, K., Fukushima, T.: Mining product reputations on the web. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 341–349 (2002)
2. Huangfu, Y., et al.: An improved sentiment analysis algorithm for Chinese news. In: 12th International Conference on Fuzzy Systems and Knowledge Discovery, pp. 1366–1371 (2015)
3. Nasukawa, T., Yi, J.: Sentiment analysis: capturing favorability using natural language processing. In: Proceedings of the KCAP-03, 2nd International Conference on Knowledge Capture, pp. 70–77 (2003)
4. Dave, K., Lawrence, S., Pennock, D.M.: Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In: Proceedings of International Conference on World Wide Web, pp. 519–528 (2003)
5. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of 2nd Conference on Empirical Methods in Natural Language Processing, vol. 10, pp. 79–86 (2002)
6. Das, S., Chen, M.: Yahoo! for amazon: extracting market sentiment from stock message boards. In: Proceedings of APFA, pp. 1375–1388 (2001)
7. Tong, R.M.: An operational system for detecting and tracking opinions in on-line discussion. In: Proceedings of SIGIR Workshop on Operational Text Classification (2001)
8. Turney, P.D.: Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of Annual Meeting of the Association for Computational Linguistics, pp. 417–424 (2002)
9. Wiebe, J.: Learning subjective adjectives from corpora. In: Proceedings of National Conf. on Artificial Intelligence, pp. 735–740 (2000)
10. Vidya, N.A., Fanani, M.I.: Budi I: Twitter sentiment to analyse net brand reputation of mobile phone providers. Procedia Comput. Sci. **72**, 519–526 (2015)
11. Karlgren, J., Cutting, D.: Recognizing text genres with simple metrics using discriminant analysis. In: Proceedings of COLIN, pp. 1071–1075 (1994)
12. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Foundation and trends in Information Retrieval, vol. 2, no. 1–2, pp. 1–135 (2008)
13. Annett, M., Kondrak, G.: A comparison of sentiment analysis techniques: polarizing movie blogs. In: proceedings of 21st Canadian Conference on Artificial Intelligence, pp. 25–35 (2008)
14. Mahmood, T., et al.: Mining Twitter big data to predict 2013 Pakistan election winner. In: proceedings of 16th International Multi Topic Conference (INMIC), pp. 49–54 (2013)
15. Saif, H., He, Y., Alani, H.: Semantic sentiment analysis of Twitter. In: Proceedings of 11th International Conference on Semantic Web, vol. 1 part 1, pp. 508–524 (2012)

16. Mukherjee, S., Bhattacharyya, P.: Sentiment analysis in Twitter with lightweight discourse analysis. In: Proceedings of 19th International Conference of Database Systems for Advanced Applications, pp. 1847–1864 (2013)
17. Jiang, L., Yu, M., et al.: Target dependent Twitter sentiment classification. In: Proceedings of the 49 h Annual Meeting for Computational Linguistics, pp. 151–160 (2011)
18. Chawla, K., et al.: IIT-B sentiment analysts: participation in sentiment analysis in Twitter SemEval 2013 task. In: Proceedings of Annual Meeting of the Association of Computational Linguistics (2013)
19. Joshi, A., Baramurali, A.R., Bhattacharyya, P.A.: Fall back strategy for sentiment analysis in Hindi: a case study. In: Proceedings of the 49th International Conference on NLP, pp. 105–112 (2010)
20. Yang, X.S.: A meta heuristic Bat inspired algorithm. Nat. Inspir. Co-op. Strateg. Optim. Stud. Comput. Intell. **284**, 65–74 (2010)
21. Ramesh, B., Mohan, V.C.J., Reddy, V.C.: V: Application of bat algorithm for combined economic load and emission dispatch. Int. J. Electr. Eng. Telecommun. **2**(1), 1–9 (2013)
22. Hatzivassiloglou, V., Wiebe, J.M.: Effects of adjective orientation and gradability on sentence subjectivity. In: Proceedings of COLING, pp. 299–305 (2000)
23. Gandomi, A.H., Yang, X.S., Talatahari, S., Deb, S.: Coupled eagle strategy and differential evolution for unconstrained and constrained global optimization. Comput. Math. Appl. **63**(1), 191–200 (2012)
24. Yang, X.S.: Bat algorithm: literature review and applications. Int. J. Bio Inspir. Comput. **5**(3), 141–149 (2013)
25. Mishra, S.X., Shaw, K., Mishra, D.: A new Meta-heuristic Bat Inspired Classification Approach for Microarray Data. Procedia Technology, pp. 802–806 (2012)
26. Nakamura, R.Y.M., Yang, X.S., et al.: BBA: a binary bat algorithm for feature selection. In: 25th SIBGRAPI Conference of Graphics, Patterns and Images, pp. 291–297 (2012)
27. O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A.: From tweets to polls: linking text sentiment to public opinion. Time Series Fourth International AAAI conference on weblogs and social media (2010)
28. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Predicting elections with twitter: what 140 characters reveal about political sentiment. In: Fourth International AAAI Conference on Weblogs and Social Media (2010)
29. Billsus, D., Pazzani, M.J.: A hybrid user model for news story classification. In: proceedings of the Seventh International Conference on User Modeling, pp. 99–108 (1999)
30. Go, A., Bhayani, R., Huang, L.: Twitter Sentiment Classification Using Distant Supervision. Stanford University Press (2009)
31. Mukherjee, S., Bhattacharyya, P.: Feature specific sentiment analysis for product reviews. Part 1, Lect. Notes Comput. Sci. 475–487 (2012)
32. Mayo, M.: A Clustering Analysis of Tweet Length and its Relation to Sentiment. Computation and Language Journal. Cornell University Press, Ithaca (2014)
33. He, Yulan, Zhou, Deyu: Self training from labelled features for sentiment analysis. Inf. Process. Manage. **47**(4), 606–616 (2011)
34. Leong, C.H., Lee, Y.H., Mak, W.K.: Mining Sentiment in SMS texts for Teaching Evaluation. Expert Systems with Applications, pp. 2584–2589 (2012)
35. Verma, S., Bhattacharyya, P.: Incorporating semantic knowledge for sentiment analysis. In: Proceedings of 6th International Conference on Natural Language Processing, pp. 1–6 (2008)
36. Aggarwal, A., Xie, B., Vovsha, I., et al.: Sentiment Analysis of Twitter Data. In: proceedings of ACL Workshop on Languages in Social Media, pages 30–38 (2011)

Internet of Things: A Survey on IoT Protocol Standards

Karthikeyan Ponnusamy and Narendran Rajagopalan

Abstract In the coming years, resource constrained devices which can be uniquely identified as an object and communicate seamlessly, often called as the Internet of Things (IoT) needs a standard technology to evolve in this real world. The emerging IoT technology will lead to the advancement in various fields like smart city, communication, healthcare, industry, transportation, security, education, research work and environmental services. And several industry bodies and standard forums are researching to develop a protocol which will satisfy all the special requirements and security, needed by constrained devices with limited processing power and resources. This paper surveys some of the main IoT protocol standards with its limitations, and solutions are proposed for future research works.

Keywords Internet of Things (IoT) · XMPP · AMQP · DSS · MQTT MQTT-SN · CoAP · UDP · DTLS

1 Introduction

Internet of Things (IoT) which is also sometimes referred as Web of Things has not been around for a long time. All the smart devices are intelligently communicating with each other and interacting with the environment due to the huge availability of IPv6 address space which leads to IoT [1]. IoT devices can be classified into two types namely, resourceful devices and constrained devices. Resourceful devices have high-processing power and resources whereas, constrained devices have limited processing power and resources. Most of the smart applications are running in constrained devices which have limited resources. These devices have to sense

K. Ponnusamy (✉) · N. Rajagopalan
Department of Computer Science, NIT Puducherry, Karaikal, India
e-mail: karthi8356@gmail.com

N. Rajagopalan
e-mail: narenraj@gmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_64

the environment, communicate with other devices and act according to their analyzed results automatically.

To make the communication between the IoT devices simple and secure, a standard protocol is required. The protocols will have different system architectures like client-server, publish subscriber, peer-to-peer, bus, and tree or star structure according to their needs. The protocol should be designed in such a way that, scalability of nodes should not affect the performance of the protocol, in low-power and lossy network (LLN) the connection between the nodes may fail often which has to be handled dynamically by the protocol, handle resource requirements at low-cost, provide interoperability and security. This survey is focused on an overview of all the main IoT protocol standards currently available with their performance and issues in the real world. In the final section, solutions are proposed to overcome the challenges faced by the IoT protocol standards for future research works.

2 IoT Protocol Standards

The Internet communications infrastructure evolves rapidly and appropriate mechanisms are required to communicate with such devices. These improvements lead to the development in various application areas like healthcare, weather forecast, smart grid, home equipment, industries, scientific researches, and smart cities, among many others. Many IoT protocol standards are being proposed to satisfy the limitations of the constrained devices. Some of the IoT protocol standards are Extensible Messaging and Presence Protocol (XMPP), Data Distribution Service (DDS), Advanced Message Queuing Protocol (AMQP), Message Queue Telemetry Transport (MQTT), Message Queue Telemetry Transport for Sensor Networks (MQTT-SN), Constrained Application Protocol (CoAP), and so on.

2.1 Extensible Messaging and Presence Protocol (XMPP)

The open standard technology, Extensible Messaging and Presence Protocol (XMPP) provide instant messaging and scalability features for the IoT applications. It is used in many real-time applications like group chat, audio, and video calls, etc. XMPP supports asynchronous communication by exchanging the data using XML streams between the server and the client, where the XMPP addressing is used to locate them uniquely in the network. According to [2], to differentiate from the Representational State Transfer (REST) architecture which is popularly known in World Wide Web, XMPP architecture is named as Availability for Concurrent Transactions (ACT). Simultaneous transaction of the information is possible in this architecture either between the client and the server or between the servers in the available network.

In XMPP, instant messages can be sent between the devices on the Internet, independent to the operating system they are running. The data to be exchanged must be structured data called as XML stanzas. These XML stanzas are exchanged over the network between two or three nodes. XMPP runs over TCP, sometimes even it runs over HTTP which is again run over TCP. XMPP is distributed client-server architecture, therefore, only if the client connects with the server it can gain access to the network. Once the client got the access to the network, it can exchange the XML stanza with other clients which have got access to the network with its server.

XMPP has a specific way of addressing the device with the domain name (name@domain.com). XMPP uses unique addressing similar to Domain Name System (DNS), so that the messages in the network can be routed and delivered to the destination. All the XMPP clients and XMPP servers present in the network are addressable. And services that are accessed by the XMPP clients and XMPP servers are addressable. As in [3], <domain_name> is known as server addresses for example "im.example.com," every server has one or more accounts which can be denoted by <local_name> along with the <domain_name> for example, "account1@im.example.com" and every account has one or more resources which can be uniquely identified by the resource path attached along with the <domain_name> and the <local_name> for example "account1@im.example.com/resource1."

During the communication process, the client connects to the server, exchanges the XML stanzas and then closes the connection. While connecting to the server, it has to determine to which IP address and port number it has to connect. Based on the domain

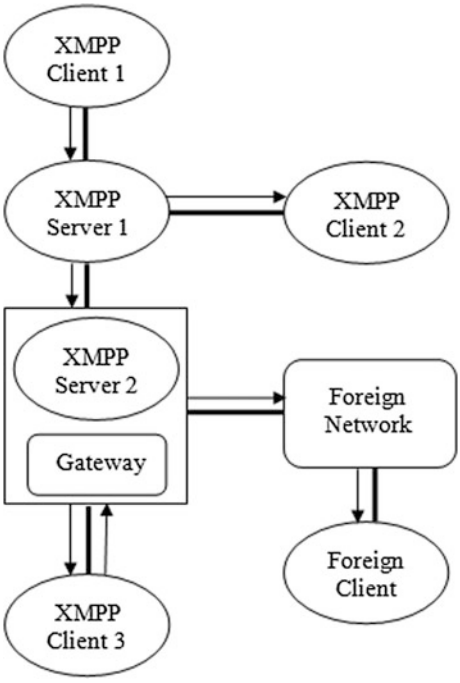


Fig. 1 XMPP architecture

name, IP address, and port number of the server to be connected can be identified. For opening the XML stream, it needs the IP address and port number since it sends XML streams over TCP and the IP address and port number can be determined while initiating. Then for secure communication Transport Layer Security [TLS] is used, which encrypts the communication channel. After binding the resources to the stream, XML stanzas can be exchanged with other entities on the network continuously. Finally, XML stream and TCP connections are closed (Fig. 1).

In XMPP, server to server communication is allowed. One XMPP server can exchange the XML stanzas with the other server on exchanging the terms between them in order to create inter-domain communication. That is instead of exchanging XML stanzas with other nodes on the network, one server can directly exchange XML stanzas continuously with another server and the server can exchange XML stanzas if the connected clients need to communicate with the other server.

XMPP provides a great way in communicating with IoT devices. The advantage of XMPP is scalability which improves the performance of IoT applications. Also, XMPP has secure authentication and encryption built into its core specifications and servers can be isolated.

2.2 Advanced Message Queuing Protocol (AMQP)

AMQP is mainly used for business purposes where they are more concerned about the messages exchanged. In organizations, the business message exchanged has to provide the necessary information and reliability on the network which can be provided by AMQP. The main goal of the AMQP is to deliver the messages without loss and provide security and interoperability. The publishers and subscribers send the messages over TCP for communication which provides the reliable point-to-point connection. Therefore, the receiver has to send an acknowledgement for every received messages (Fig. 2).

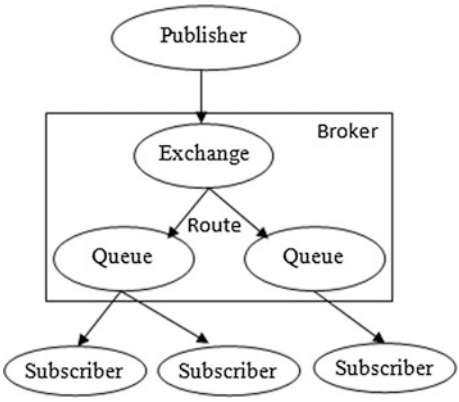


Fig. 2 AMQP architecture

In AMQP system, producer generates the message and consumer processes the message generated by the producer. Queues are entities existing inside the broker, which is used to store and forward messages. The broker will receive the messages from the producer and store it in separate queues and then it is forwarded to the receivers. The messages are routed by predefined rules and conditions. The security for AMQP system is provided by the Transport Layer Security (TLS). The protocol definitions for security layers are defined in the AMQP specification. As described in [4], the SASL security layer depends on its host protocol to provide secure communication.

The AMQP is used in enterprise messaging applications as described in [5], since it requires high performance, throughput, scalability, and reliability. It is also used to communicate with other applications on different systems. It is most appropriate for the operations need to be performed on the server.

2.3 Data Distribution Service (DDS)

DDS [6] is a data communications standard developed by the Object Management Group (OMG). It is used in distributed applications for its low-latency data communications. In DDS, no need of configuring any gateway or server since it is a peer-to-peer model and it supports communication between different DDS implementations. Because of the data-centric middleware, it is mainly used in industry and embedded applications where high performance is needed (Fig. 3).

It has a flexible publish subscribe architecture which will lead to a loose coupling between data architecture. The flexible nature of the architecture makes the DDS systems to change dynamically, according to the requirements. DDS supports dynamic discovery and the exchange of data between DDS nodes such as

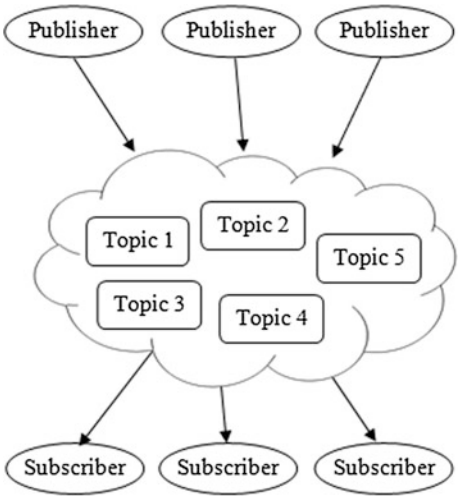


Fig. 3 DDS architecture

publishers, subscribers, database and other additional services by using publish subscribe model [7]. It provides Dynamic Discovery of publishers and subscribers. So, there is no need for the IoT applications to identify the information about the endpoints for communication as DDS helps to discover the publishers and subscribers automatically. It also provides type safety by providing the permission to the applications, to define the data types, which are used by the application during the communication [7]. The application then reads and writes according to the specified data type.

The Quality of Service (QoS) policies of DDS can be set according to the need of the implementation, such as reliability, security, etc. High-performance device systems use DDS as it provides flexibility, reliability, security, and speed, which are required in building complex real-time applications.

2.4 Message Queue Telemetry Transport (MQTT)

Message Queue Telemetry Transport (MQTT) is a simple and a lightweight messaging protocol which runs on top of the TCP and uses publish subscribe model. MQTT is an open standard technology and it is developed by IBM [8]. MQTT purpose is to connect the embedded devices over the network to communicate with the limited use of resources. MQTT helps in reducing the resource requirements for the constrained resource devices by providing the efficient use of bandwidth, reliability and different levels of QoS. MQTT focus on large networks where many constrained devices are connected and needed to be managed by the server.

Publisher, subscriber, and broker are the three components of MQTT. One or more clients can connect to a broker and subscribe the topics to which notifications are required. Clients can also publish a topic by connecting with the broker. Many clients may subscribe to the same topics. According to [8], a series of packets are exchanged in MQTT protocol in a defined way known as MQTT Control Packets. An MQTT Control Packet consists of a fixed header, variable header, and payload. The message contains a fixed header in the first two bytes. In this format, the value of the Message Type field indicates a variety of messages, such as CONNECT, CONNACK, PUBLISH, PUBACK, PUBREC, PUBREL, PUBCOMP, SUBSCRIBE, SUBACK, UNSUBSCRIBE, UNSUBACK, PINGREQ, PINGRESP, and DISCONNECT. The DUP flag indicates that the message is the duplicate delivery of a PUBLISH control packet. The QoS flag and RETAIN flag are to represent the PUBLISH quality of service and retain flag, respectively (Fig. 4).

MQTT defines three levels of QoS deliver once, at least once and exactly once. The QoS defines how the broker and the client must communicate to ensure the delivery of a message. The sender can send the message using any QoS level and while clients subscribing the topics it can use any QoS level. Therefore, the client will choose the maximum level of QoS it received. If the level of QoS has increased then the reliability also increases which leads to high latency and bandwidth requirements [8]. QoS level 0 is to deliver the message once where no confirmation

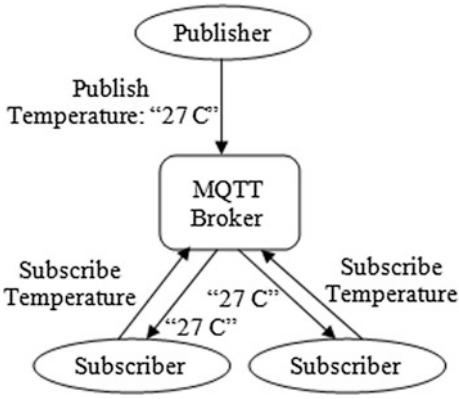


Fig. 4 MQTT architecture

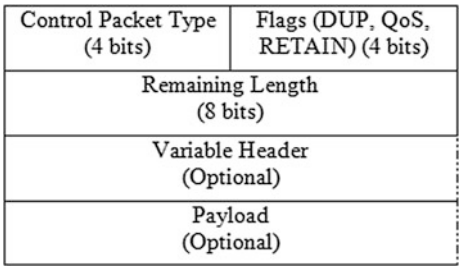


Fig. 5 MQTT message format

is needed. QoS level 1 is to deliver the message at least once, where confirmation is required. QoS level 2 is to deliver the message exactly once where four step handshake is required. If a message is set to be retained then the broker will keep the message even after sending the message to all the current subscribers. If a client subscribes a topic, where the messages are marked as retained then the client will receive all the messages instantly. If a topic is not updated frequently and there is no retained message then the client which is recently subscribed has to wait for a long time to receive a notification. If the retained message is available then the client will get notification instantly. All messages can be set to retained message (Fig. 5).

The variable header part of the Control Packet type contains a two bytes Packet Identifier field. PUBLISH, PUBACK, PUBREC, PUBREL, PUBCOMP, SUBSCRIBE, SUBACK, UNSUBSCRIBE, and UNSUBACK are some of the Control Packet types. The Client and Server assign Packet Identifiers which is unrelated to each other. The client should send the CONNECT control packet to the server after the network connection is established. The payload is optional part and it contains one or more encoded fields. As described in [8], it contains unique Client identifier for the client, user name, password, with topic and with message. All the fields are optional except the client identifier and it can be identified from the flags present in the variable header part. Sometimes username and password have to be provided by the clients to connect for authentication. The TCP connection is encrypted with SSL/TLS to provide secure communication for MQTT protocol.

2.5 MQTT for Sensor Networks (MQTT-SN)

MQTT-SN is a publish/subscribe messaging protocol similar to MQTT protocol, for wireless sensor networks (WSN), and made suitable for Sensor and Actuator solutions which have limited resources and processing power, as in [9]. Since MQTT uses TCP to connect with the broker, it has to keep an open connection with the broker for a long time. And the topic string used is too long then it will be violating the rules of 802.15.4. So, MQTT-SN supports UDP to communicate and topic id is used to identify the resources. Topic id is created by indexing the topic of two-byte long. MQTT clients, gateways, and forwarders are three components used in MQTT-SN. If the gateway is not integrated with the broker then MQTT protocol is used to communicate between the gateway and the broker. There are two methods of the standalone gateway, transparent gateway and aggregated gateway [9]. In transparent gateway method, every client will have end-to-end connection with the broker via gateway where MQTT-SN will be mapped to MQTT. It is simple and easy to implement. In aggregated gateway method, only one MQTT connection is used between the broker and the clients via the gateway. It is difficult to implement but in the case of more number of clients, it will reduce the overload.

MQTT-SN message contains two parts, header part which has the size 2 or 4 bytes and optional variable part. The header part contains message length and message type. The message length may be of 1 or 3 bytes, which is used to represent the size of the message. The message type is 1 byte, which is used to represent one of the many message types, such as PUBLISH, SUBSCRIBE, CONNECT, DISCONNECT, REGISTER, and so on [9]. The variable part contains fields such as client id, data, duration, flags, add gateway, gateway id, message id, protocol id, radius, return code, topic id, topic name, will message, will topic. These fields of the variable part are used to configure the messages of different types. The Quality of Service (QoS) levels followed in MQTT-SN are similar to MQTT protocol. And it is also optimized for the IoT devices which operate on battery and have limited resource and processing power.

2.6 Constrained Application Protocol (CoAP)

CoAP is an open standard and primarily designed for resource constrained devices by Internet Engineering Task Force (IETF). CoAP [10] is based on the Representational State Transfer (REST) architecture and it runs over User Datagram Protocol (UDP) [11]. Since it uses UDP, to make it reliable the messages are marked as "confirmable" or "non-confirmable."

If the message is marked as confirmable then the receiver should send the acknowledgment and if the message is marked as non-confirmable then it does not require acknowledgment. An acknowledgment message will be sent only if the

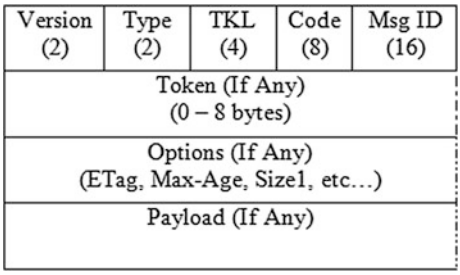


Fig. 6 CoAP message format

confirmable message is received and it may also carry data which is known as a piggybacked response. Reset message is sent by the receiver if the received message is corrupted or if the receiver node got rebooted and not able to interpret the message. If the server cannot able to respond immediately to a received confirmable message then, in that case, it sends an acknowledgment without any content, so that the client will stop retransmitting the message. Later, if the server sends the respond message separately, then it is known as a separate response. As CoAP runs over UDP, Datagram Transport Layer Security (DTLS) [12] is used to provide security (Fig. 6).

CoAP message has fixed header part of 4-byte size, token, option, and payload parts, as described in [10]. These headers, token, and options of the CoAP messages are binary encoded to reduce the protocol overhead. Message header part contains Version (2 bits) to represent version of CoAP, Type (2 bits) to represent the message type (confirmable, non-confirmable, acknowledgement, or reset), Token Length (4 bits) to represent the size of the Token, Code (8 bits) to represent the status code and Message ID (16 bits) to match acknowledgment/reset with confirmable/non-confirmable messages and used to identify the duplicate message.

The header part is followed by the Token field which varies from 0 to 8 bytes. The token is generated by the client in such a way that the source/destination endpoint pair is unique. Therefore, the token value should be echoed in the response message. The token part is followed by option part which contains ContentType, Etag, LocationPath, LocationQuery, MaxAge, ProxyUri, ProxyScheme, UriHost, UriPath, UriPort, UriQuery, Accept, IfMatch, IfNoneMatch, and Size. These options can be classified under two categories either critical or elective [10]. Critical means the endpoint should understand the proper decoding and Elective means option could be ignored by the endpoint. These are defined to handle the options which are not understandable by the process. The option part is followed by the Payload which can be included by both requests and responses, depending on the Method or Response Code, respectively.

CoAP endpoints have the capability of caching the responses. Therefore, previous responses can be reused to respond another similar request in future and by doing so, it reduces the response time and network bandwidth consumption. A proxy [10] is the CoAP endpoint used to forward the request and performs caching. In forward proxy endpoint is selected by the client and in Reverse proxy

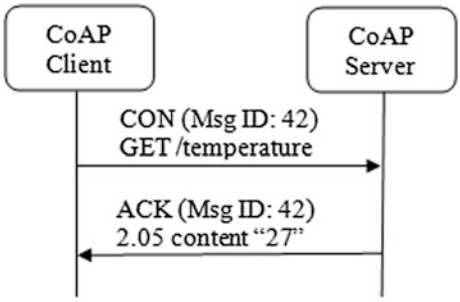


Fig. 7 Confirmable message

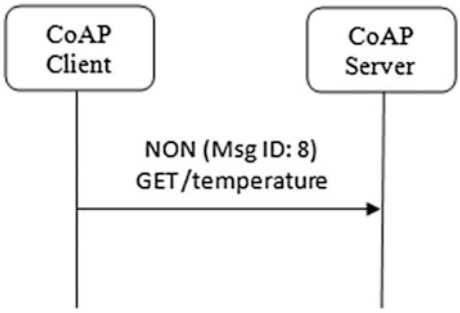


Fig. 8 Non-confirmable message

stands in for origin servers. Cross-Proxy is used to translate between different protocols. There are two ways to access a reverse via forward proxy, CoAP-HTTP, and HTTP-CoAP proxying. Similar to HTTP, CoAP also uses GET, POST, PUT, and DELETE request methods. GET method is used to retrieve the information corresponding to the resource in request URI. POST method is used to request the processing of representation in the response. PUT method is used to update or create the resource identified by the request URI. DELETE method is used to delete the resource identified by URI in the request message. The response code (8 bit) [10] is split into two parts 3-bit and 5-bit class, it is represented in "c.dd" format. If the response code is of 2.xx then it is Success response. If the response code is of 4.xx then it is Client Error. If the respond code is of 5.xx then it is Server Error (Figs. 7 and 8).

Discovering the services in CoAP is of two types, service discovery and Resource discovery. In service discovery, clients learn from the URI of a the resource that references the resource present in the server or by multicasting to find the server address and in resource discovery, servers provide a list of their available resources at the Uri path/well-known/core and using this link, the client can discover the resources provided by the server. The DTLS security for CoAP [10] provides four types of security mode. In NoSec mode, the DTLS is disabled and alternative security is provided in lower layer security. In the PreShared Key mode, DTLS is enabled and the node has a list of keys, where each key is used to communicate with a node. In RawPublicKey mode, DTLS will be enabled and the

device will contain an asymmetric key pair with the list of nodes to which the device can communicate. In Certificate mode, DTLS will be enabled and the device will contain an asymmetric key pair along with an X.509 certificate which is bonded with its subject and it is signed by some common trust root.

CoAP has a special feature in addition to HTTP known as Observe resource. It is added to the CoAP request model with the ability to observe a resource which is similar to subscribe option in publish/subscribe model. As in [13], it provides multicast support, reliability, low overhead, security, and simplicity for constrained environments.

3 Iot Standards Limitations and Proposed Solutions

These IoT protocols have been developed to meet the requirements of constrained devices which have limited amount of memory and computing. But each protocol has its own advantages and disadvantages. The advantages of XMPP are addressing, speed and scalability. The drawback of XMPP is the lack of end-to-end encryption. In many cases the encryption may not be necessary, but in future most of the IoT devices will eventually need it. The lack of end-to-end encryption is a major drawback for IoT manufacturers. Role-based access control technology and providing a secure token, XMPP security can be improved. AMQP is an open standard with publish/subscribe architecture. Its advantage is to track all the messages and ensure that each message is delivered successfully as intended, in spite of failures. DDS is a flexible publish subscribe architecture developed by the Object Management Group (OMG). Its advantages are flexibility, reliability, and speed necessary to build complex, real-time applications. But both AMQP and DDS are mainly used by high-performance integrated device systems to build complex, real-time applications. A new message format can be proposed to communicate with the IoT device, so that IoT devices need not require high performance. In AMQP, the broker can map the new message format to the XMPP message format.

The MQTT is a simple and a lightweight messaging protocol which runs over TCP and uses publish subscribe model. Since MQTT uses TCP, every MQTT client has to retain a connection open to the broker at all times. Also, the resource topic names often have lengthy strings which makes them practically impossible for 802.15.4. To overcome these limitations MQTT-SN is defined. MQTT-SN runs over UDP and additional indexing feature for the resource topics are added to the broker. But, however, only few protocols are supported and the pub-sub mechanism is complex. CoAP is a RESTful protocol which runs over UDP. The CoAP messages are binary encoded, which reduces the protocol overhead. But the disadvantage is the lack of appropriate key management mechanisms for the support of secure CoAP multicast communications. As described in [14], Group key management mechanism can be implemented by exchanging a session key between the devices involved in group communication during DTLS handshake. And end-to-end security is not

provided while mapping techniques is used to integrate TLS and DTLS. As described in [15], TCP/TLS or UDP/DTLS methods can be used based on the network conditions dynamically to ensure the end-to-end security.

4 Conclusion

In the rapidly evolving IoT technology, a standard IoT protocol is needed which solves almost all the complexities faced by the constrained devices. Many IoT protocols are being proposed, and each protocol focuses on different goals such as reliability, security, speed, and so on. For analysis, it is appropriate to use AMQP. MQTT is the best choice to use in data communication, where the device collects the data. XMPP is commonly used to connect devices with the people, so it can be used where the message is used to control or communicate with a device. DDS is used to connect the devices and the distributed applications. CoAP is a good choice to use when the devices in the network are operating on battery, or the processing power is limited. In this paper, all the important IoT protocol standards are discussed and solutions are proposed to improve the efficiency of the protocols. For future research works, this survey of IoT protocol standards and the proposed solutions will be helpful to design or choose an efficient IoT protocol standard for specific applications.

References

1. Ashton, K.: That 'Internet of Things' thing. RFID J. http://www.rfidjournal.com/articles/view?4986 (2009)
2. Saint-Andre, P.: Extensible messaging and presence protocol (XMPP): core. In: Internet Engineering Task Force (IETF), [RFC6120], March 2011
3. Saint-Andre, P.: Extensible messaging and presence protocol (XMPP): address format. In: Internet Engineering Task Force (IETF), [RFC6122], March 2011
4. OASIS Standard: OASIS Advanced Message Queuing Protocol (AMQP) Version 1.0 (2012)
5. AMQP working group protocol specification: AMQP: Advanced Message Queuing, version 0.8. http://www.iona.com/opensource/amqp/amqp0-8-june19.pdf (2006)
6. Data Distribution Services Specification, Version 1.2. http://www.omg.org/spec/DDS/1.2/ (2007)
7. Esposito, C., Russo, S., Di Crescenzo, D.: Performance assessment of OMG compliant data distribution middleware. In: Parallel and Distributed Processing, 2008, IEEE International Symposium, pp. 1–8 (2008)
8. OASIS Standard: Message Queue Telemetry Transport (MQTT) Version 3.1.1. http://docs.oasis-open.org/mqtt/mqtt/v3.1.1/os/mqtt-v3.1.1-os.pdf (2014)
9. Stanford-Clark, A., Truong, H.L.: MQTT for Sensor Networks (MQTT-SN) Protocol Specification Version 1.2. International Business Machines Corporation (IBM) (2013)
10. Shelby, Z., Hartke, K., Bormann, C.: The constrained application protocol (CoAP). In: Internet Engineering Task Force (IETF), [RFC7252], June 2014
11. Postel, J.: User datagram protocol (UDP). In: Internet Engineering Task Force (IETF), [RFC0768], August 1980

12. Rescorla, E., Modadugu, N.: Datagram transport layer security (DTLS). In: Internet Engineering Task Force (IETF), [RFC6347], January 2012
13. Bormann, C., Castellani, A.P., Shelby, Z.: CoAP: an application protocol for billions of tiny internet nodes. IEEE Internet Comput. **16**, 62–67 (2012)
14. Granjal, J., Monteiro, E., Sá, Jorge: Security for the internet of things: a survey of existing protocols and open research issues. IEEE Commun. Surv. Tutorials **17**, 1294–1312 (2015)
15. Brachmann, M., Garcia-Morchon, O., Kirsche, M.: Security for practical CoAP applications: issues and solution approaches. GI/ITG KuVS Fachgesprch Sensornetze (FGSN). Universitt Stuttgart (2011)

Influence of Twitter on Prediction of Election Results

Prabhsimran Singh and Ravinder Singh Sawhney

Abstract Twitter is a social networking website (SNW), where people post their likes and feelings related to a particular issue. Over the past few years, it has become a tool for people worldwide to express their political sentiments regarding their favorite political party. This has opened a new domain for researchers to develop certain techniques those can be used to predict the outcome of an election based on the political sentiment of the persons through the tweet posted by them. This paper discusses the previous work done in this field by various researchers on political outcome of elections held in different countries and try to establish most stable and appropriate technique in predicting the election results.

Keywords Twitter · Tweet · Election prediction · Sentiment analysis Opinion mining

1 Introduction

Twitter was launched in 2006 as a social networking website (SNW), where people can post anything up to a limit of 140 words. Earlier, it was used just to express one's thoughts but slowly people realized that it can be utilized as a powerful tool to express their feelings and sentiments towards an entity. This field of study that analyzes one's opinion, sentiment toward an entity is called sentiment analysis [1]. With the increase in popularity of twitter, sentiment analysis has been applied on a large number of applications. From past few years, researchers are using sentiment analysis to find the political sentiments of people toward an individual or a party

P. Singh (✉)
Department of Computer Science, Guru Nanak Dev University, Amritsar, India
e-mail: prabh_singh32@yahoo.com

R. S. Sawhney
Department of Electronics Technology, Guru Nanak Dev University, Amritsar, India
e-mail: sawhney.ece@gndu.ac.in

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_65

going to contest the upcoming elections. This paper concentrates on political predictions through tweets using sentiment analysis.

This paper evaluates the previous work done in this field, studies their features, merits, and demerits. In addition to these, some other factors (Gray Areas) will be discussed that lead to successful prediction or failure of each proposed technique/model. Through this study we will try to answer two main questions:

Q1: *Is any technique/model proposed till date is efficient enough to predict election outcome of any country just on basis of tweets?*
Q2: *What are the major challenges faced in these predictions?*

We would discuss these two issues during the flow of this paper. First, all the techniques proposed till date in this field are discussed. Next, these techniques are compared based on the various parameters and analysis of their result is done. Finally, we conclude our paper with a possible suggestion which can improve future work.

2 Related Work/Literature Survey

The election prediction using sentiment analysis is a relatively new area of research so a limited amount of work has been done till date. Now we will briefly discuss all related work done in this field.

Tumasjan et al. [2], predicted the outcome of German Federal election held on 27th September, 2009. They collected 104,003 tweets from 13th August to 9th September, 2009 of six popular political parties. Their technique was simply based on the number of tweets that a party gets involving the name of the party or its prominent leaders. At the end, the mean absolute error was calculated. The result predicted by their proposed technique was correct. A major reason behind the correct prediction was that Germany is a developed country [3]. Moreover, the number of people having internet connections covers almost 88% of the total German population [4]. With such a large population contributing, correct results were obvious.

Jungherr et al. [5], used a similar approach to support vector machines (SVM) embedded with Wahlgetwitter to predict the outcome of German Federal elections held in 2009. 10,085,982 Tweets were collected from 18th June to 1st October, 2009. They collected tweets that have hashtags containing only political mention, in this way they covered all the parties contesting the election. However, the result predicted by them was incorrect as according to them the winner should have been Pirate Party but the party failed to win even a single seat. So they concluded that the party with the highest number of tweets may not necessarily get large vote share in the elections.

Sang and Bos [6], worked to predict the number of seats that a party can win in Dutch Senate Election held in March 2011. They collected 64,395 tweets in 8 days prior to the elections and counted the number of tweets received by a party and

multiplied it with sentiment weight which was simply a number of positive tweets upon the total number of tweets. The prediction of seats was only true in case of two parties out of total 12 parties, hence the method was found to be not so accurate. One major factor that played against this technique was that the time period of tweet collection was very small, i.e., 8 days, hence, the prediction was not up to the mark.

Choy et al. [7], collected 16,616, tweets from 17th to 25th August, 2011 to predict the outcome of 2011 Singapore Presidential Elections. The main highlight of this paper was that not only they counted positive votes for each of the four candidates contesting the election, they also characterized the tweets according to different age groups. It was clear the older population was in favor of PAP party. However, the actual prediction of the winner was incorrect. Again as discussed above the major factor that went against their prediction was a small time period of tweet collection.

Makazhanov and Rafiei [8], stated that along with the number of tweets number of re-tweets and positive replies should also be taken into account. They collected 181,972 tweet samples for 2012 Alberta (Canada) elections and further applied methods like J48 (Decision Tree Based), Naïve Bayes (NB) and Sentistrength (S), etc. and compared them. They also classified the users into active and silent users and correctly predicted the outcome of elections from the tweets of silent users. Being a developed country [3] and 85% of the population having access to the internet [4] were major reasons that led to correct predictions in their case.

Ceron et al. [9], used HK model Proposed by Hopkins and King to predict the 2012 Presidential Election of USA and Italy. Moreover, MAE was used to calculate error from actual vote share received by the winning candidate. They stressed that not every age group, gender or group is equally involved in social media. Further, they predicted that only counting the tweets and mentioning the candidate name or party name is not a good way to provide accurate prediction. This technique correctly predicted the outcome of both elections.

Nooralahzadeh et al. [10], applied their technique to predict the outcome of 2012 presidential elections held in two developed countries, i.e., USA and France [3]. The basic principle behind their technique was scoring function and number of positive tweets received by each candidate. At the end prediction made for both countries were correct. Unlike other, they counted only positive tweets from the users. Populations of USA, French having access to the internet are 88% and 89%, respectively [4], played an important role in correct prediction of the results.

Gaurav et al. [11], said that for precise predictions tweet of each twitter user should only consider once, this helps to eliminate the phenomenon of Twitter Bomb, i.e., multiple tweets coming from a single user. They collected around 400 million tweets for each of the three countries, i.e., Paraguay, Ecuador, and Venezuela for 2013 elections. They used two techniques for election predictions Moving Average Aggregate Probability (MAPP) and Moving Average Aggregate Probability Using Counts (MAPC). Their results correctly predicted the outcome of all three countries. High literacy rate, i.e., greater than 94% in all the three countries [12], became a major factor toward correct prediction of the results.

Mahmood et al. [13], collected tweets using a twimemachine website and they categorized the tweets into Pro (in favor) and Anti (against) for each of the political party. Then they applied Decision Tree, Naïve Bayes, and Support Vector Machine (SVM) on these tweets to predict the results of 2013 Pakistan general elections. All these techniques predicted that Pakistan Tehreek-e-Insaf (PTI) party will emerge as winner, however, the actual elections were won by Pakistan Muslim League [Nawaz] (PMLN) party, hence, the technique was not successful. The major reasons that led to this wrong prediction were the fact that Pakistan is a developing country [3], moreover, literacy rate is just 57% [12].

Almatrafi et al. [14], classified the tweets into positive and negative tweets. One main point in that study was the geo-location of the tweet, i.e., the location from where the tweet was sent. They applied Naïve Bayes on 650,000 tweets of two popular political parties AAP and BJP to predict the outcome of 2014 Indian general election. They also claimed that during election period political and social events gives sharp rise to a number of tweets. The results showed that BJP will emerge victorious as it was more popular on twitter, which was correct as compared to actual prediction.

Prasetyo and Hauff [15], stated using twitter for predicting the results of elections in developing countries is more accurate and precise method. They performed their study on 2014 Indonesia's Presidential Elections and compared their results to 20 offline methods of poll prediction where twitter outperformed all other predictions. They also stated factors like location, keyword selection that can influence the results of precise predictions.

Srivastava et al. [16], proposed a mapping function that converts tweet share into seat share. They collected positive tweets for three main parties AAP, BJP, and Congress contesting 2015 Delhi Assembly Elections. The positive tweets showed that AAP will be the Victorious party as it got more than 54% tweet share. Further using their proposed function they accurately predicted the seat share of each party. Delhi having a high literacy rate of 86% leads to correct predictions by them.

3 Comparative Study

This section compares all the techniques/method discussed in the previous section based upon various factors like country, type of election, number of tweets, tweet collection period, method used, literacy rate, and prediction results. Table 1, shows computed results of comparison based on these factors. The factor Type of Election is classified as following:

A = *In this type of voting system people are directly voting for a candidate which is standing in Election for the post of Head of State/Country. This system is followed in countries like USA, France,* etc.

Table 1 Comparative analysis of different approaches

Paper references	Country under consideration	Type of election	Number of tweets	Tweet collection period (days)	Technique/method used	Literacy rate (%)	Prediction outcome
[2]	Germany	B	104,003	36	Counting number of favorable tweets + mean absolute error (MAE)	99	Correct
[5]	Germany	B	10,085,982	102	Support vector machine (SVM) + Wahlgetwitter	99	Incorrect
[6]	Holland	B	64,395	8	Number of tweets + sentiment weight	98	Incorrect
[7]	Singapore	A	16,616	8	Counting number of positive tweets	96	Incorrect
[8]	Alberta (Canada)	B	181,972	–	J48 (decision tree based), Naïve Bayes (NB) and Sentistrength (S)	99	Correct
[9]	USA	A	50,000,000	40	HK method + mean absolute error (MAE)	99	Correct
[9]	Italy	A	50,000	50	HK method + mean absolute error (MAE)	99	Correct
[10]	USA	A	196,000	60	Scoring method + number of positive tweets	99	Correct
[10]	France	A	10,000	10	Scoring method + number of positive tweets	98	Correct
[11]	Paraguay	A	397,000,000	6	MAPP and MAPC	95	Correct
[11]	Venezuela	A	400,000,000	6	MAPP and MAPC	95	Correct
[11]	Ecuador	A	395,000,000	6	MAPP and MAPC	94	Correct
[13]	Pakistan	B	–	130	Decision tree, naïve Bayes and support vector machine (SVM)	58	Incorrect
[14]	India	B	650,000	5	Naïve Bayes	72	Correct
[15]	Indonesia	A	7,020,228	130	Counting number of favorable tweets + mean absolute error (MAE)	93	Correct
[16]	Delhi (India)	B	352,730	31	Positive sentiment share (PSS)	86	Correct

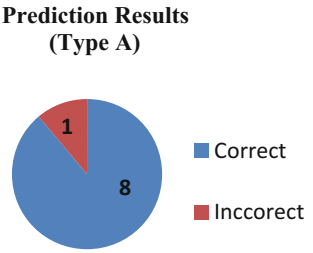


Fig. 1 Prediction results (type A)

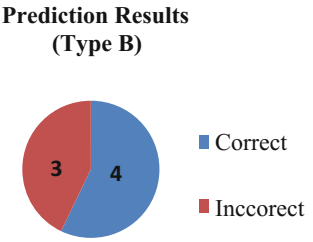


Fig. 2 Prediction results (type B)

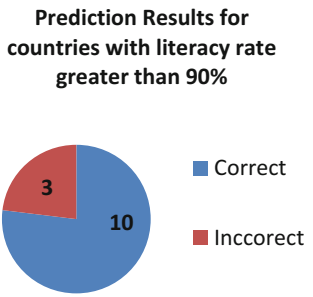


Fig. 3 Prediction results for countries with literacy rate greater than 90%

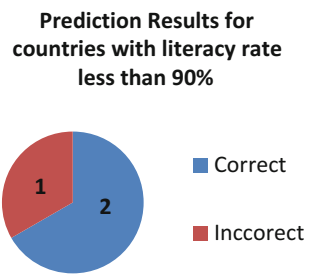


Fig. 4 Prediction results for countries with literacy rate less than 90%

B = *In this type of voting system people vote for respective candidates of different parties standing from their region/constituency,* etc. *Then the party with most number of winning candidates selects one candidate as head of State/Country. This system is followed by Countries like India, Pakistan,* etc.

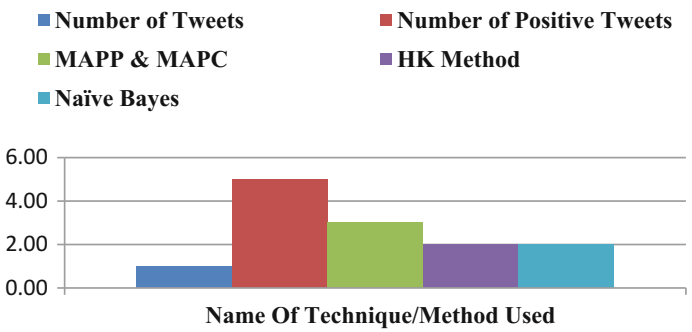


Fig. 5 Techniques which gave true predictions

Figure 1, shows the result of Prediction outcomes based on "Type of Elections," where ***Type A*** election got 8 out of 9 results correct predictions, i.e., success rate of 88%. While Fig. 2, shows that only 4 out of 7 cases in ***Type B*** got correct predictions, i.e., success rate of 57%. Figure 3, shows the countries with Literacy rate more than 90% got 10 out of 13 predictions correct, i.e., success rate of 76%. While Fig. 4, shows only 2 out of 3 predictions were correct for countries with Literacy rate less than 90% achieving a success rate of only 66%. Figure 5, shows the result of Techniques that gave true predictions. It is clear that counting number of positive tweets give true prediction in maximum number of cases, i.e., 5 times.

4 Conclusion and Future Work

This paper explored the newest field of Sentiment Analysis, i.e., to predict election outcome using tweets. The aim of the paper was to get answers of two basic questions. The answer to the ***Q1*** is no, although counting number of positive tweets gave a maximum number of correct prediction but it failed to produce 100% success rate. From the study of various papers, we came across various factors that influence the prediction results and hence answer our ***Q2***. Major factors that influence the prediction of election results are:

(a) Number of population that uses Twitter
(b) Twitter users that are actual voters
(c) Literacy Rate of the Country
(d) Percentage of population having access to Internet
(e) Geographical location of the Tweet sender
(f) Concept of Twitter bomb should be avoided
(g) Re-Tweets and emoticons should also be taken into account for predictions

(h) Time span for tweet collection
(i) Type of elections (discussed in Sect. 3)
(j) Type of keyword selection

All the above factors influence the preciseness and accuracy of prediction in one way or the other. These factors should be kept in mind so that a universal technique could be derived or proposed which can accurately predict the election of any country/region. Moreover, for elections of type B (discussed in Sect. 3) an equally important system should be developed that can convert the predicted tweet share into seat share, as the winner in this type of election is decided on basis of party having a maximum number of seats.

References

1. Liu, B.: Sentiment analysis and opinion mining. Synth. Lect. Hum. Lang. Technol. **5**(1), 1–167 (2012)
2. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Predicting elections with twitter: what 140 characters reveal about political sentiment. In: ICWSM, vol. 10, pp. 178–185 (2010)
3. International Monetary Fund. http://www.imf.org/external/pubs/ft/weo/2015/01/weodata/groups.html
4. Central Intelligence Agency. https://www.cia.gov/library/publications/resources/the-world-factbook/rankorder/2153rank.html
5. Jungherr, A.: Tweets and votes, a special relationship: The 2009 federal election in germany. In: Proceedings of the 2nd Workshop on Politics, Elections and Data, pp. 5–14. ACM (2013)
6. Sang, E.T.K., Bos, J.: Predicting the 2011 Dutch senate election results with twitter. In: Proceedings of the Workshop on Semantic Analysis in Social Media, pp. 53–60. Association for Computational Linguistics (2012)
7. Choy, M., Cheong, M.L.F., Laik, M.N., Shung, K.P.: A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census correction. http://arXiv.org/abs/1108.5520 (2011)
8. Makazhanov, A., Rafiei, D.: Predicting political preference of Twitter users. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 298–305. ACM (2013)
9. Ceron, A., Curini, L., Iacus, S.M.: Using sentiment analysis to monitor electoral campaigns: method matters—evidence from the United States and Italy. Soc. Sci. Comput. Rev. 0894439314521983 (2014)
10. Nooralahzadeh, F., Arunachalam, V., Chiru, C.: 2012 Presidential elections on twitter–an analysis of how the US and French election were reflected in tweets. In: 2013 19th International Conference on Control Systems and Computer Science (CSCS), pp. 240–246. IEEE (2013)
11. Gaurav, M., Srivastava, A., Kumar, A., Miller, S.: Leveraging candidate popularity on twitter to predict election outcome. In Proceedings of the 7th Workshop on Social Network Mining and Analysis, p. 7. ACM (2013)
12. Central Intelligence Agency. https://www.cia.gov/library/publications/resources/the-world-factbook/fields/2103.html
13. Mahmood, T., Iqbal, T., Amin, F., Lohanna, W., Mustafa, A.: Mining twitter big data to predict 2013 Pakistan election winner. In: Multi Topic Conference (INMIC), 2013 16th International, pp. 49–54. IEEE (2013)

14. Almatrafi, O., Parack, S., Chavan, B.: Application of location-based sentiment analysis using Twitter for identifying trends towards Indian general elections 2014. In: Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication, p. 41. ACM (2015)
15. Dwi Prasetyo, N., Hauff, C.: Twitter-based election prediction in the developing world. In: Proceedings of the 26th ACM Conference on Hypertext and Social Media, pp. 149–158. ACM (2015)
16. Srivastava, R., Kumar, H., Bhatia, M.P., Jain, S.: Analyzing Delhi assembly election 2015 using textual content of social network. In: Proceedings of the Sixth International Conference on Computer and Communication Technology 2015, pp. 78–85. ACM (2015)

The Rise of Internet of Things (IoT) in Big Healthcare Data: Review and Open Research Issues

Zainab Alansari, Safeeullah Soomro, Mohammad Riyaz Belgaum and Shahaboddin Shamshirband

Abstract Health is one of the sustainable development areas in all of the countries. Internet of Things has a variety of use in this sector which was not studied yet. The aim of this research is to prioritize IoT usage in the healthcare sector to achieve sustainable development. The study is an applied descriptive research according to data collection. As per the research methodology which is FAHP, it is a single cross-sectional survey research. After data collection, the agreed paired comparison matrices, allocated to weighted criteria and the priority of IoT usage were determined. Based on the research findings, the two criteria of "Economic Prosperity" and "Quality of Life" achieved the highest priority for IoT sustainable development in the healthcare sector. Moreover, the top priorities for IoT in the area of health, according to the usage, were identified as "Ultraviolet Radiation," "Dental Health," and "Fall Detection."

Keywords Internet of Things (IoT) · Healthcare · Fuzzy-AHP · Big data

Z. Alansari (✉) · S. Soomro · M. R. Belgaum
College of Computer Studies, AMA International University,
Salmabad, Kingdom of Bahrain
e-mail: zeinab@amaiu.edu.bh

S. Soomro
e-mail: s.soomro@amaiu.edu.bh

M. R. Belgaum
e-mail: bmdriyaz@amaiu.edu.bh

Z. Alansari · S. Shamshirband
College of Computer Science and Information Technology,
University of Malaya, Kuala Lumpur, Malaysia
e-mail: shamshirband@um.edu.my

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_66

1 Introduction

Internet revolution in the recent decades showed that the new kinds of technologies can affect all aspects of the businesses [1]. Nowadays, with the help of new technologies such as network connections, wireless communication, and sensors, ubiquitous communication is always possible. Moreover, Business owners reflected many articles regarding IoT and introduced it as new solutions in ICT which they believe it has the potential to earn a great and valuable income.

The purpose of IoT is empowering the objects for connection at anytime, anywhere, with anything and any person that gets the ideal use of any route, service or network. Internet of Things is a new evolution of the Internet. IoT is a new technology that focuses on the environmental effect and deals with a variety of wireless and wired connections which are communicating with each other. These objects are working together to develop an application for a new service and achieve a common goal together. In fact, it considers as a development challenge for creating a great smart world. A world which in its actual shape is digital and virtual but moves toward the development of smart world which creates smarter areas of energy, transportation, health, cities, and many more [2, 3].

In practice, different countries have numerous motivations to support IoT including United States [4], China [5], European Union [6] and India [7]. According to the report from IoT European Research Cluster (IERC), three motivations for the development of IoT in countries is Economic Prosperity, Quality of Life, and Environmental Protection [8]. It has been discussed the sustainable development literature [9].

No priority of IoT has been determined in the health sector, and it seems to be essential to prioritize some areas of IoT which have the highest potential for sustainable development in the health sector. Furthermore, the use of innovative technologies has always been considered by researchers, but no research has been conducted on IoT in the health sector so far. The aim of this study is to prioritize the functional areas of the health sector which is the development of IoT to achieve sustainability in the health sector. Therefore, the study seeks to answer the following questions:

- How much is each indicator's weight of Economic Prosperity, Quality of Life and Environmental Protection, to assess the IoT in the health sector?
- What is the priority of each IoT application in the health sector?

2 Literature Review

The Internet of Things for the first time was used in 1999 which described the world where anything, including people, animals, plants, and even inanimate objects (such as cars), being able to have their digital identity which is allowing the computers to organize and manage them.

Nowadays the internet is connecting all the people, but the IoT is connecting all the objects together, and the people can control and manage them by the use of available applications in smartphones and tablets [10]. Actually, IoT is a new concept in the technology and communication world which considered as a modern technology provides the capability of sending data for anything (human, animal, or object) via network connection rather Internet or an intranet [11].

Businesses have focused heavily on IoT, and this has led to the Electronic Business (e-Business) development [12] and in many cases, customer relationship management is easier through IoT [13]. In fact, IoT is an approach that will improve the interoperability between an object with object, object with human, and human with an object and with the help of such an approach the new services will appear [14]. Moreover, one of the primary objectives of IoT is to increase intelligence in life, business, and economy [15].

2.1 Smart Health

IoT promises market potential in the field of e-health services and the telecommunications industry [16]. IoT can enhance business intelligence in hospitals and ease serving the patients in health sectors [17]. The Internet of Things can improve the health grounds and prevent diseases by providing ongoing monitoring activities to ordinary people or prone patients [16, 18]. Furthermore, it empowers the patients and helps the businesses to profit from this new innovative in the market. Somehow, it improves the patient's Social problems, people's concerns about health and the quality of their lives [19] also contributes to economic prosperity in health sector [20]. With the help of this technology the hospital activities impact on the environment (Such as production and elimination of hospital waste) can be better managed and less likely to damage the environment [21].

The usages of IoT can develop some platforms which provide smart and innovative services to patients and people in need of medical attention. Furthermore, improves their health, Security, ease of access to emergency medical care, continuing care, and quick support also improving the quality of life [16].

2.2 Applicable Areas of IoT in the Health Sector

IoT European Research Cluster (IERC), has been presented a comprehensive classification of relevant areas of IoT in smart health in the health sector. Some usages are of the type of services, and some are a kind of product. Related Areas of IoT in The Health Sector (smart health) are:

- Fall Detection: This usage focused on helping the physically challenged and elderly people in their lives so that they can live independently;
- Medical Fridges (Internal temperature protective control): Some organic elements must be kept in containers with certain conditions (temperature). IoT can well assume this task and Cause objects interaction;
- Sportsmen Care: The application used to measure the Weight, sleep, exercise, blood pressure, and other relevant parameters for professional athletes;
- Patient Surveillance: used for remote in-hospital monitoring, (especially the elderly) or used for patient's home care;
- Chronic Disease Management: Taking care of patients with chronic diseases while there is no need of physical attendance. This technology reduces the presence of people in hospitals and results in lower costs, reduces hospital stay and reduces traffic (Even reduces fuel consumption);
- Ultraviolet Radiation: UV rays Measurement and notifying the people not to enter certain areas or refrain of exposure to UV rays at certain hours;
- Hygienic Hand Control: By linking devices, such as designed RFID for emissions measurement, environmental pollution could be identified;
- Sleep Control: Devices that by linking to individuals, identifies some signs, such as heart rate, blood pressure during sleep and the data may be collected and will be analyzed after;
- Dental Health: Bluetooth-enabled Toothbrush with the help of smartphone apps records someone's brushing information to study the person's brushing habits and share the statistics with the dentist [16].

2.3 IoT Sustainability Indicators in Health Sector

United Nations [22] described the sustainability concept that seeks to define human needs without compromising the ability of future generations to meet their needs. On the other hand, Porter and Kramer [23] stated that businesses are the primary cause of the social, environment, and economy problems in recent years. Lack of business confidence causes the political leaders to develop policies that weaken competitiveness and economic growth. Selection of such policies over the last few decades created the impression that the "Economic Prosperity" is a reverse of "Social progress".

The government must learn the management in such way to ease the creation of shared social and economic value, rather than to stop it. Creating simultaneously shared value focuses on the relationship between social and economic progress and has the power of encouraging the next wave of global growth [24]. Porter and Kramer [23] stated that the idea of creating shared value clarifies the government's role in responsible behavior toward the sustainable development.

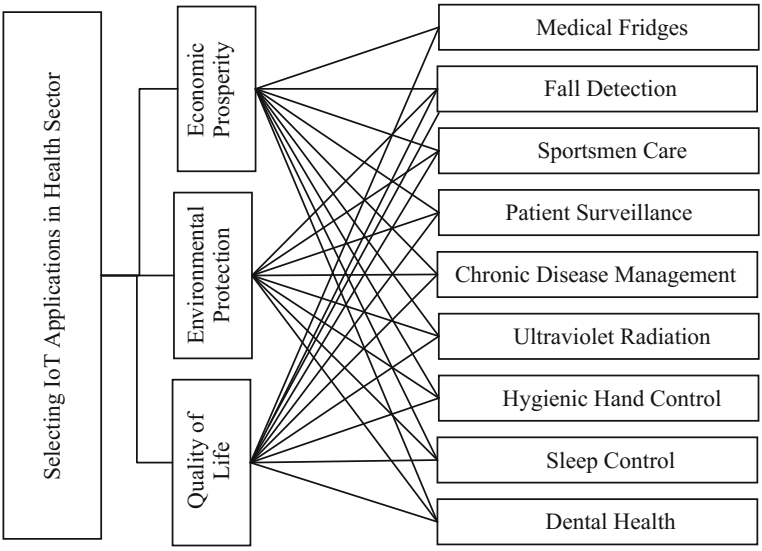


Fig. 1 The conceptual model of the study based on FAHP

In this study, Economic Prosperity, Quality of Life and Environmental Protection criteria has been used to assess the IoT scope in the health sector. As previously mentioned, IERC introduced this criterion as drivers for the IoT development [25], and it is considered in sustainable development literature as well [26]. In the research literature, criteria of economic prosperity were placed in the financial domain [27] Quality of life were set in the social sphere [28], and environmental protection included in an environmental field [29].

3 Conceptual Model

Figure 1 shows the Conceptual Model of the study which prioritized IoT in health sector according to sustainable development criteria and based on the Fuzzy Analytical Hierarchy Process (FAHP) method:

4 Methodology

The study objective is an applied research while in terms of data gathering it is considered as a descriptive nonexperimental among the quantitative researches. Since Fuzzy Analytic Hierarchy Process (FAHP) is used for weighting and

prioritization of IoT in the health sector (Smart Health), it is a Cross-sectional study among the descriptive studies.

Initially, each of the economic prosperity, quality of life and environmental protection criteria were weighted. Then in three pair-wise comparison questionnaire, each application of IoT in the health sector was compared separately according to their criteria. Finally, the Decision Matrix was obtained. Content validity method used to assess the validity of questionnaire and 12 experts in IoT Health Sector confirmed the decision-making matrix components such as criteria and options. The statistical population of the study consisted of experts who are familiar with IoT that has a background of trade cooperation or was business partners to provide services or advice to hospitals and health centers in the use of new technologies.

Due to the limited number of experts, the snowball method was used; Therefore, after referring to the Communication Research Centre, Experts in the health field were identified as a pioneer of IoT research objects. A joint meeting was handled with IoT experts to reach an agreement on each of the paired comparisons. Twenty experts have invited which only 12 experts attended the meeting and the obtained pair-wise comparison questionnaire were the base on data analysis of this study.

4.1 Fuzzy Analytical Hierarchy Process (FAHP)

AHP model is a Multi-Attribute Decision-Making (MADM) that professor saaty proposed in 1970 [30]. In 1996, a Chinese researcher named Chang proposed a fuzzy-AHP method based on the development analysis [31]. In fact, FAHP method was developed based on AHP and fuzzy logic [32]. Triangular Fuzzy Numbers are the numbers used in this method. The geometric space in such a fuzzy environment is shown in Fig. 2.

In this method, the membership function and fuzzy scale defined in Table 1.

The stages of applying FAHP method in this research are as follows:

1. Evaluation of literature research to determine the sustainability criteria and the use of IoT applications in health sector (smart health);
2. Formation of a decision team to examine the questionnaire validity;

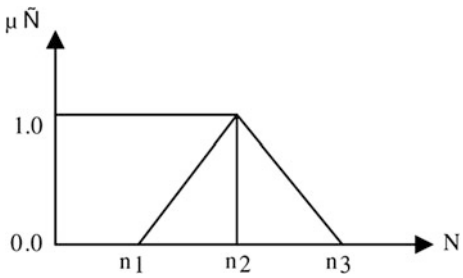


Fig. 2 A triangular fuzzy number, N [32]

Table 1 Membership functions and the definition of fuzzy scale [32]

Intensity of importance	Fuzzy number	Definition	Membership function
9	$\tilde{9}$	Extremely more importance (EMI)	(8, 9, 10)
7	$\tilde{7}$	Very strong importance (VSI)	(6, 7, 8)
5	$\tilde{5}$	Strong importance (SI)	(4, 5, 6)
3	$\tilde{3}$	Moderate importance (MI)	(2, 3, 4)
1	$\tilde{1}$	Equal importance (EI)	(1, 1, 2)

3. Distributing the questionnaires and creating paired comparisons matrix with Fuzzy terminology for each questionnaire;
4. Weighting of criteria and determining the scores of the options for each of the criteria using FAHP;
5. Ranking the usage of IoT in the health sector by FAHP method.

The calculation of FAHP method to achieve the weight of criteria and prioritization of IoT applications in the health sector conducted with Chang [31] method.

5 Analysis and Results

Determining the weights of criterion and the score of each were completed after conducting several meetings to fill the agreed questionnaire. Four paired comparison questionnaire (one questionnaire to compare paired criterion and three questionnaires to compare the IoT applications in health sector according to each criterion) with the help of fuzzy words (Table 1). Using the FAHP, criterion's weight obtained from the paired comparison agreed on matrices as shown in Fig. 3.

Similarly, each option in each of the criteria calculated using FAHP, and decision matrix was formed as shown in Table 2.

5.1 Ranking the Options Using FAHP

Ultimately, with the help of weights obtained for sustainable development criteria (Fig. 3) and decision matrix (Table 2), The final score and rank of each IoT application in health sector achieved using the FAHP as shown in Table 3.

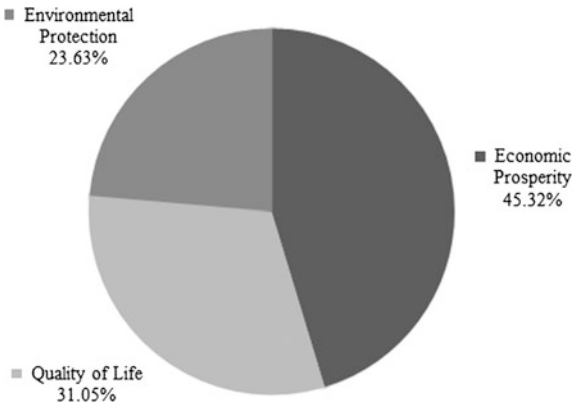


Fig. 3 Weight of criteria sustainability

Table 2 Decision matrix derived from the average weight of each criterion for each option

Options	Economic prosperity	Quality of life	Environmental protection
Fall detection	0.109	0.096	0.14
Medical fridges	0.084	0.11	0.039
Sportsmen care	0.069	0.0836	0.153
Patient surveillance	0.117	0.0954	0.121
Chronic disease management	0.079	0.104	0.025
Ultraviolet radiation	0.193	0.132	0.176
Hygienic hand control	0.098	0.094	0.12
Sleep control	0.068	0.143	0.059
Dental health	0.183	0.142	0.167

Table 3 Scores and the priority of IoT in the health sector

Options	Rank	Score
Ultraviolet radiation	1	0.0567
Dental health	2	0.0555
Fall detection	3	0.0374
Patient surveillance	4	0.0371
Hygienic hand control	5	0.0340
Sportsmen care	6	0.0311
Sleep control	7	0.0297
Medical fridges	8	0.0271
Chronic disease management	9	0.0247

6 Conclusions

According to the results, the most important criteria of IoT in the health sector for sustainable development is economic prosperity with a weight of 45.32% and then the quality of life with the weight of 31.05%. According to the Experts, the weight of Environmental Protection to develop IoT smart health sector is 23.63%. Therefore, it is recommended to the policymakers of the health sector to focus on developing the new technologies such as IoT application on economic criteria like employment and revenue, then social criteria, such as increasing the welfare of patients and citizens and satisfaction of hospital personnel in the use of medical tools. Moreover, the environmental impact of these technologies in the environmental sector, such as probable radiation, harmful radio waves, preventing the creation of waste and water should not be forgotten.

Based on Table 2, If only the criteria of Economic Prosperity are concerned in IoT field, the "Ultraviolet Radiation," "Dental Health," and "Patient Surveillance" are in priority. According to the same results, if only the criteria of Quality of Life is considered, "Sleep Control," "Dental Health," and "Ultraviolet Radiation" are in top priority and improve the citizen's quality of life more than another criterion. Finally, if only the criteria of Environmental Protection are considered, "Ultraviolet Radiation," "Dental Health," and "Sportsmen Care" are in top priorities more than any other are of IoT in the health sector or smart health contribute to the improvement of environmental protection.

However, in this study, all the three criterion are effective based on their importance in the IoT health sector. The ranking results based on FAHP of the expert's opinion and the basis for paired comparisons of IoT applications in health sector and according to the triple sustainability criteria which were obtained in this research (Table 3), indicate that the priority of IoT in health sector using sustainable development criteria are "Ultraviolet Radiation," "Dental Health," "Fall Detection," "Patient Surveillance," "Hygienic Hand Control" and "Sportsmen Care". "Sleep Control," "Medical Fridges," and "Chronic Disease Management" are located in the last rank. Therefore, it is recommended to the government and relevant health centers to give more support to the IoT application in "Ultraviolet Radiation," "Dental Health," and "Fall Detection" as they followed the most interests of stability.

7 Future Studies

The results of this study increase the knowledge of IoT, the familiarity of IoT innovation in the health sector, and encourages the usage of new technologies according to the sustainable development criterion. However, this study has some limitations, since the IoT applications experience in the health sector is limited, it may influence the motivation of investment based on the given prioritization which

seems the need for technical and economic feasibility studies. Moreover, the governments are still not supporting the IoT governance, its regulation, and consumer and producer's rights. Therefore, it could be a future studies concern of IoT implementation.

References

1. Premkumar, G., Roberts, M.: Adoption of new information technologies in rural small businesses. Omega **27**(4), 467–484 (1999)
2. Atzori, L., Iera, A., Morabito, G.: The internet of things: a survey. Comput. Netw. 2787–805 (2010)
3. Sundmaeker, H., Guillemin, P., Friess, P., Woelfflé, S.: Vision and challenges for realising the Internet of Things. In: CERP- IoT—Cluster of European Research Projects on the Internet of Things (2010)
4. U.S. Government Promoting Development Of The "Internet of Things". http://smartamerica.org/news/u-s-government-promoting-development-of-the-internet-of-things
5. China Internet of Things. http://iot.cqna.gov.cn
6. IERC (European Research Cluster on the Internet of Things). http://www.internet-of-things-research.eu/about_iot.htm
7. Deity (Department of Electronics & Information Technology). http://deity.gov.in/content/draft-internet-thingsiot-policy
8. Smith, I.G.: The internet of things 2012 new horizons. In: European Research Cluster on the Internet of Things, UK (2012)
9. Carter, C.R., Easton, P.L.: Sustainable supply chain management: evolution and future directions. Int. J. Phys. Distrib. Logist. Manage. **41**(1), 46–62 (2011)
10. Ashton, K.: That 'internet of things' thing. RFiD J. **22**(7), 97–114 (2009)
11. Chui, M., Löffler, M., Roberts, R.: The internet of things. McKinsey Q. **2**, 1–9 (2010)
12. Uckelmann, D., Harrison, M., Michahelles, F.: Architecting the Internet of Things. Springer (2011)
13. Xiaocong, Q., Jidong, Z.: Study on the structure of "Internet of Things (IOT)" business operation support platform. In: Communication Technology (ICCT), 12th IEEE International Conference, pp. 1068–1071 (2010)
14. Miorandi, D., Sicari, S., DePellegrini, F., Chlamtac, I.: Internet of things: vision, applications and research challenges. Ad Hoc Netw. **10**(7), 1497–1516 (2012)
15. Tan, L., Wang, N.: Future internet: The internet of things. In: 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE), vol. 5, p. 376. IEEE (2010)
16. Vermesan, O., Friess, P.: Internet of Things-From Research and Innovation to Market Deployment. River Publishers (2014)
17. Roman, R., Najera, P., Lopez, J.: Securing the internet of things. Computer **44**(9), 51–58 (2011)
18. Bandyopadhyay, D., Sen, J.: Internet of things: applications and challenges in technology and standardization. Wirel. Pers. Commun. **58**(1), 49–69 (2011)
19. Helal, A., Cook, D.J., Schmalz, M.: Smart home-based health platform for behavioral monitoring and alteration of diabetes patients. J. Diabetes Sci. Technol. **3**(1), 141–148 (2009)
20. Haller, S., Karnouskos, S., Schroth, C.: The Internet of Things in an Enterprise Context, pp. 14–28. Springer, Berlin (2009)
21. Perera, C., Zaslavsky, A., Christen, P., Georgakopoulos, D.: Context aware computing for the internet of things: a survey. In: Communications Surveys & Tutorials, vol. 16.1, pp: 414–454. IEEE (2014)

22. United Nations (UN): UN Documents: gathering a body of global agreements. http://www.un-documents.net/wced-ocf.htm
23. Porter, M.E., Kramer, M.R.: Creating shared value. Harvard Bus. Rev. **89**(1/2), 62–77 (2011)
24. Crane, A., Palazzo, G., Spence, L.J., Matten, D.: Contesting the value of "creating shared value". Calif. Manag. Rev. **56**(2), 130–153 (2014)
25. Smith, I.G.: The internet of things 2012 new horizons. In: European Research Cluster on the Internet of Things, Halifax, UK (2012)
26. Carter, C.R., Easton, P.L.: Sustainable supply chain management: evolution and future directions. Int. J. Phys. Distrib. Logist. Manage. **41**(1), 46–62 (2011)
27. Bansal, P.: Evolving sustainably: a longitudinal study of corporate sustainable development. Strateg. Manage. J. **26**(3), 197–218 (2005)
28. Baud, I.S.A., Grafakos, S., Hordijk, M., Post, J.: Quality of life and alliances in solid waste management. In: contributions to urban sustainable development. Cities **18**(1), 3–12 (2001)
29. Zhang, K.M., Wen, Z.G.: Review and challenges of policies of environmental protection and sustainable development in China. J. Environ. Manage. **88**(4), 1249–1261 (2008)
30. Saaty, T.L.: What is the Analytic Hierarchy Process?, pp. 109–121. Springer, Berlin (1988)
31. Chang, D.Y.: Applications of the extent analysis method on fuzzy AHP. Eur. J. Oper. Res. **95**(3), 649–655 (1996)
32. Büyüközkan, G., Çifçi, G., Güleryüz, S.: Strategic analysis of healthcare service quality using fuzzy AHP methodology. Expert Syst. Appl. **38**(8), 9407–9424 (2011)

Implementation of SSVEP Technology to Develop Assistive Devices

Manjot Kaur and Birinder Singh

Abstract With the advancements in science and technology, many assistive devices have been developed to support physically disabled individuals. Steady-state visual evoked potential (SSVEP)-based brain–computer interface is an emerging technology in this area. Work has been done to develop SSVEP-BCI based wheelchair asynchronously which moves in all the four directions. SSVEP frequencies chosen to generate brain signals are 6, 7.5, 12, and 14 Hz. Brain signals are acquired using Electroencephalogram (EEG) technique. MindMedia provides us a platform to capture and to display the signals on the screen in the time domain with amplitudes in microvolts. Signals are processed in Matlab to extract the SSVEP frequencies by considering frequency characteristic of signals. After identification of SSVEP frequency, a command corresponding to this is generated using the digital logic of 0 and 1. In a hardware implementation, a microcontroller board, driver circuit, and DC motors are used. This system is validated by ten subjects with multiple attempts in all the four directions. Results have shown high accuracy of 91.75% in the system.

Keywords SSVEP · BCI · EEG · Digital logic · Asynchronous control

1 Introduction

BCI is a technology which is predominantly meant for people suffering from locked-in syndrome. This allows people suffering from the syndrome, where they are cognitively fit but locked in a paralyzed body, with an alternative and augmentative way to interact with the environment. Using this technology, instead of using natural peripheral nerves and muscles for communication, one communicates

M. Kaur (✉) · B. Singh
Computer Science Department, BBSBEC, Fatehgarh Sahib, Punjab, India
e-mail: manjot793@gmail.com

B. Singh
e-mail: birinder.singh@bbsbec.ac.in

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_67

with the help of brain signals. These brain signals are processed in order to identify user's intent [1].

Signals generated in the brain are captured using noninvasive technique known as EEG. It measures the signals from the brain using electrodes placed on the scalp. These electrodes are made up of conductive material which senses the potential generated in the brain. EEG system measuring brain signals is shown in Fig. 3a.

The electrodes are placed on the scalp on the basis of International 10–20 system as shown in Fig. 1. This is the standard system of placing the electrodes on the head. In this system along with the active electrodes, two more electrodes are taken known as a reference electrode and the ground electrode.

The measured signal is actually the potential difference between the active electrode and the reference electrode. This is done to enhance the clarity of signal by eliminating or reducing the noise from the main signal. Reference electrode must be placed close to the brain. The ground electrode can be placed anywhere on the body and is used for rejection of unwanted signals such as noise. Reference electrode along with ground electrode [3] is shown in Fig. 2.

In order to trigger the signals in the brain, a SSVEP [4] technique is employed. SSVEP is a technique in which screen would flicker at a particular frequency between dark and light colors. A person attends to it by concentrating on the screen for some seconds without any blink [5]. Electric potential elicits in the brain in response to it whose frequency would be equal to the frequency of attended flickering stimuli. These signals are dominant in visual cortex region (present in the occipital lobe) of the brain [2, 6] This procedure is shown in Fig. 3.

Signals captured using EEG is processed in matlab to identify the SSVEP frequency. Signal processing includes the creation of feature vector, extracting the frequency component of the signals using FFT. On the basis of this command is generated in the form of digital logic of 0's and 1's. the command is transferred from the matlab to the hardware via serial COM port in the form of signals.

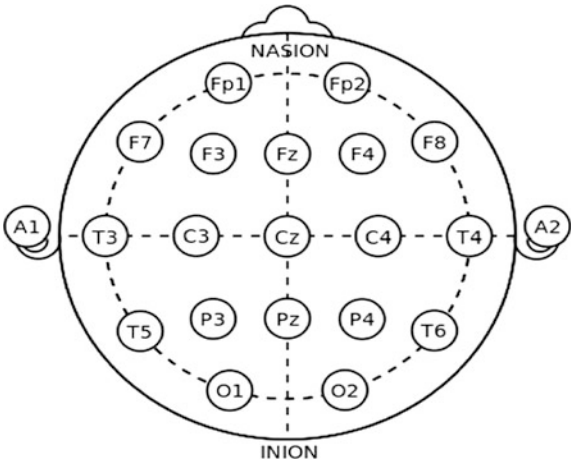


Fig. 1 International 10–20 system [2]

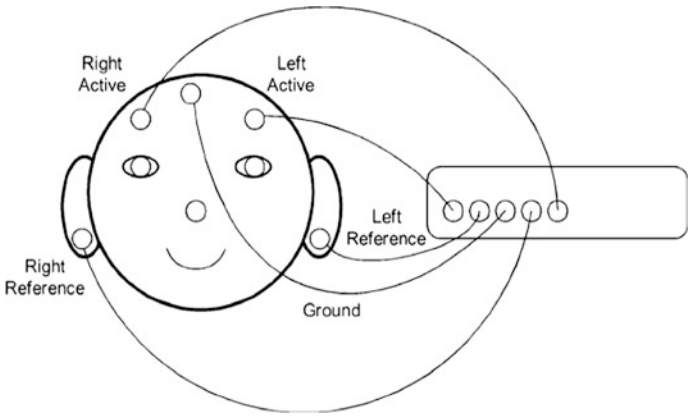


Fig. 2 Reference and ground electrode

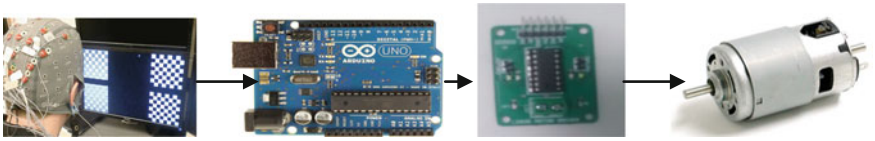


Fig. 3 **a** SSVEP technology [7], **b** Arduino uno, **c** L293D driver, **d** DC Motor

Arduino UNO microcontroller board is used to transfer the signals to the motors. It has both analog and digital pins available on it. It controls the motor using digital pins. It takes 5 V power supply from the laptop using Universal serial bus (USB) wire. Digital pins set to high when they get 5 V of supply and in case of 0's voltage drops to zero. Arduino microcontroller has been programmed using Arduino software. ATmega328 microcontroller is used to control the pins. Pins of Arduino can be used in two modes either in input mode or output mode. pinMode() function is used to settle the mode of pins and digitalWrite() function is used to set the values of the pins. Fig. 3b shows arduino uno microcontroller board.

Arduino UNO sends signals to the motor driver and motor driver further pass the same signals to the DC motors. L293D driver circuit is used which can control the operation of two motors simultaneously. It has total 16 pins but when soldered on printed circuit board it works with only six pins among which four pins A1, A2, B1 and B2 are used to receive the input from the microcontroller and two pins 5 V and ground are used for power supply purposes. Arduino supplies 5 V power to the microcontroller board. It is used because motors need a power supply of minimum 12 V but microcontroller can supply up to 5 V hence this driver circuit is used to meet the power requirements of DC motors (Fig. 3c).

Two DC motors are attached to the rear wheels of the wheelchair. These motors are small-scale motors used to develop small applications as shown in Fig. 3d. These DC motors need a power supply of 12 V provided by the motor driver

circuit. Two pins on the left and right side of motor drivers controls the operation of two motors simultaneously.

In this work, an attempt has been made to asynchronously control a wheelchair using brain signals with high accuracy. In this case, instead of using an actual wheelchair, a wheelchair prototype has been developed which is controlled using signals. Success of this work has shown that by making hardware modifications real-time wheelchair could also be controlled using the same logic.

2 Methods for Wheelchair Control

To control wheelchair in real time it needs both hardware and software support. Hence, wheelchair control has been subdivided into two parts: software implementation [8, 9] and hardware implementation.

2.1 Software Implementation

Generating Brain Signals. In order to produce brain signals Four rectangular boxes are shown on the laptop screen at top, left, bottom, and right positions with width and height of 200 * 100 pixels. These would flicker in black and white shades with unique frequencies and each frequency corresponds to a particular movement [10]. Stimuli at the top correspond to 7.5 Hz causes forward movement, stimuli at the left corresponds to 12 Hz causes left movement, Stimuli at the bottom corresponds to 6 Hz causes movement in a backward direction, Stimuli at the right corresponds to 14 Hz causes movement in right direction [11].

This is made in javascript code with each flicker interval is set in milliseconds by using following formulae given below.

$$F * N = 500. \tag{1}$$

where *F* corresponds to flickering frequency and *N* corresponds to the time interval in milliseconds after which each flicker change its shade [7, 12]. Figure 4 shows GUI for generation of brain signals.

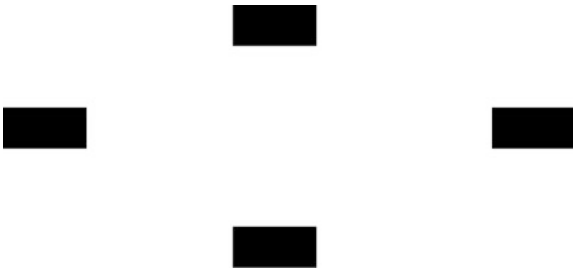


Fig. 4 GUI for generation of brain signals

Acquisition of Brain Signals. Signals generated in the brain are captured using Nexus10 Biosignal acquisition system. Electrodes made up of conductive material such as Ag/Agcl are placed on the Scalp. Electrodes are wet electrodes which are attached to the head using 10–20 gel. Since SSVEP signals are dominant in visual cortex region hence electrodes are placed in the occipital region over the point O1 or O2. The reference electrode is placed at ear mastoid and ground is placed over the forehead [12].

These signals are sent to the BioTrace+ NX10 software running on the laptop which shows the signals in the time domain as shown in Fig. 5. These signals are then converted to Excel file in order to process them in matlab as shown in Fig. 6. This file now can be loaded into matlab in CSV format for further processing.

Signal Processing. This excel file is loaded into matlab for extraction of SSVEP frequency. First of all, signals would be plotted in the time domain to analyze the variation of amplitude along with time [13] as shown in Fig. 7.

Then these signals are converted from time domain to frequency domain using fast Fourier transform (FFT). These are converted to extract frequency component of signals to extract maximum frequency [14]. The frequency spectrum for all the four frequencies is shown below. X axis contains frequency values and Y contain amplitude values [15].

After converting into frequency domain maximum frequency from this graph is extracted for classification.

Command generation and Transmission: After classification of maximum frequency digital command corresponding to it is generated in the form of bits. There are four combinations of bits used for each type of movement (Table 1).

Fig. 5 Signals in BioTrace+

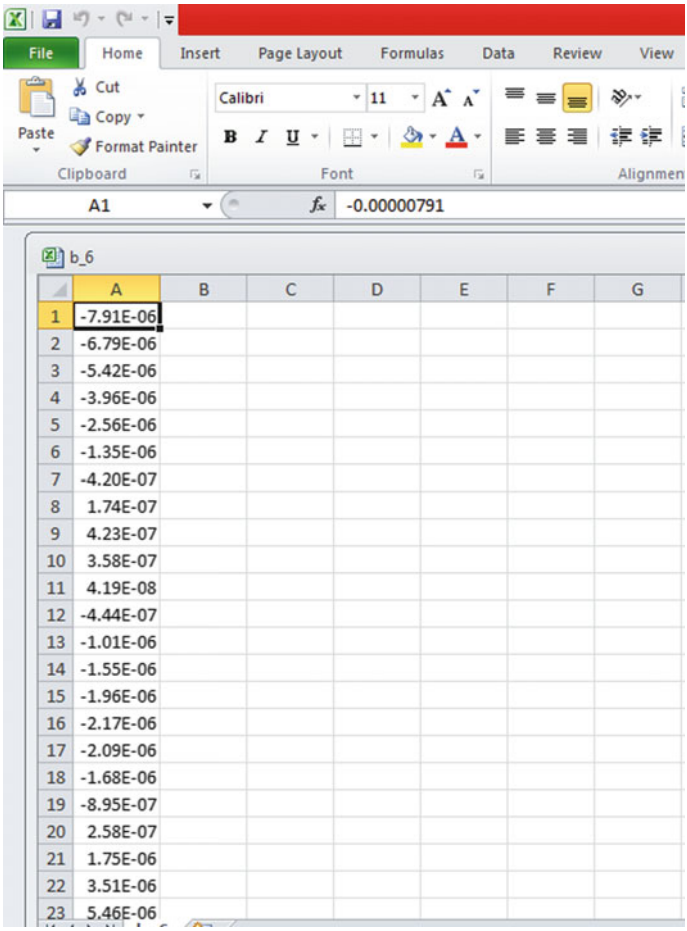


Fig. 6 Excel sheet of EEG signals

2.2 Hardware Implementation—Command Transmission Phase

Step 1: **Passing the command from Matlab to Arduino Microcontroller**. This involves sending of digital bits from the matlab to the Arduino board in the form of serial communication via COM port. Arduino IDE is used to program the microcontroller to communicate with external software such as matlab and helps us to select the port for communication and version of the board. Digital pins 2, 4, 8, and 9 are used as output pins. At 1 they receive 5 v and at 0 they receive 0 v. Arduino receives bits according to above-mentioned logic (Table 2).

Step 2: **Passage of commands from Arduino Board to L293D driver circuit**. Arduino board further sends the commands to the motor driver circuit L293D.

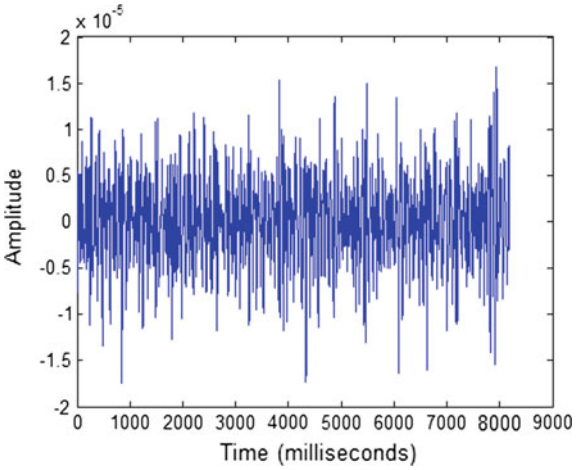


Fig. 7 EEG signals in time domain

Table 1 Table shows generated commands for each type of bits

Motor 1		Motor 2		Action
Pin 1	Pin 2	Pin 1	Pin 2	
0	1	0	1	Forward
1	0	1	0	Backward
0	0	0	1	Left
0	1	0	0	Right
0	0	0	0	Stop

Table 2 Table shows Digital logic corresponding to Arduino pins

Movement type	Pins used	Values
Forward	2, 4, 8, 9	0, 1, 0, 1
Backward	2, 4, 8, 9	1, 0, 1, 0
Left	2, 4, 8, 9	0, 0, 0, 1
Right	2, 4, 8, 9	0, 1, 0, 0
Stop	2, 4, 8, 9	0, 0, 0, 0

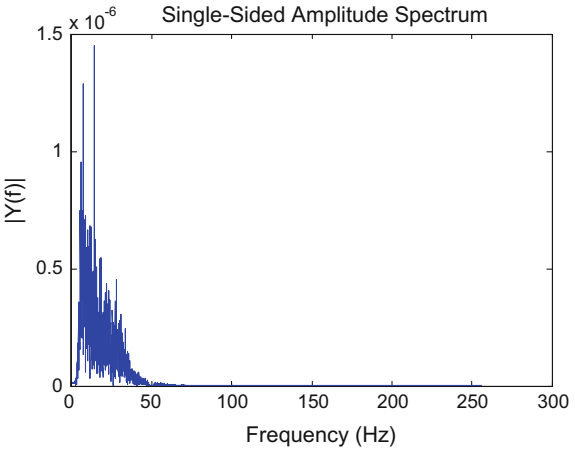
L293D has total six pins out of which two are used for power supply and rest of the four pins control the motor (Table 3).

Step 3: **Passage of commands from L293D to the DC motor**. This includes sending of commands from the motor driver to the DC motors in order to control the wheelchair. A1, A2 pins of L293D controls two pins on the left side of motor driver and B1, B2 pins controls the pins on the right side of motor driver. Pins on left of L293D controls motor1 and pins on right side of L293D controls motor 2.12 v Power supply to the motor is given by L293D which is supplied to L293D externally with the help of adapter, plugged into AC power supply. Two motors attached to the rear wheels of the wheelchair make it move with the constant speed of 100 revolutions per minute. One wheel is attached at the front to make balance

Table 3 Table shows L293D pins value corresponding to Arduino pins

Arduino pin no.	Operation	Driver pin name
2	Controls	A1
4	Controls	A2
8	Controls	B1
9	Controls	B2
5v	Controls	5v
Gnd	Controls	Gnd

Fig. 8 Right movement



among the wheels. This is the last step after which wheelchair will move in the desired direction.

3 Results

With the implementation of above methodology, a wheelchair prototype has been successfully developed and tested under captured EEG signals. Results shown that the wheelchair has been successfully controlled in any direction on the basis of frequency values. The figure shows the working model of EEG-based wheelchair (Figs. 8, 9 and 10).

3.1 Performance Evaluation

Performance of this system is evaluated using the parameter Accuracy Rate. Accuracy rate is measured in comparison to the error rate. If accuracy increases then

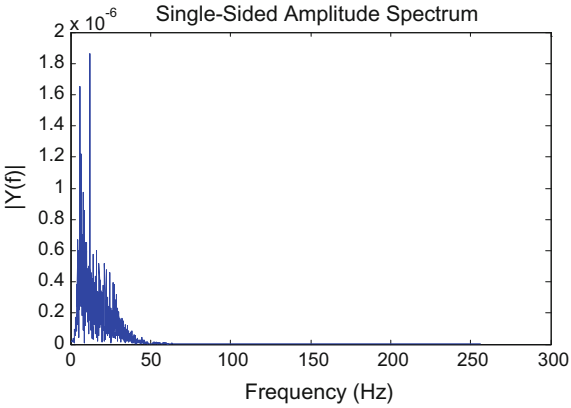


Fig. 9 Left movement

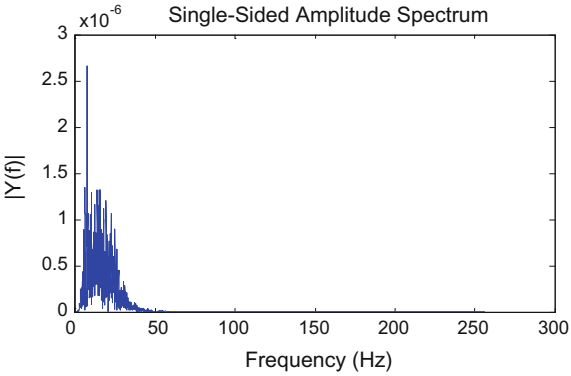


Fig. 10 Forward movement

Table 4 Table shows motion direction corresponding to frequency values

Serial no	Frequency (Hz)	Operation executed
1.	6	Backward movement
2.	7.5	Forward movement
3.	12	Left movement
4.	14	Right movement

automatically the error reduces. Error here indicates the wrong identification of the target (Table 4).

$$\text{Accuracy Rate} = \text{Correctly Identified Trials}/\text{Total number of Trials}. \quad (2)$$

$$\text{Error Rate} = \text{Incorrectly Identified Trials}/\text{Total number of Trials}. \quad (3)$$

In order to measure this ten subjects has been chosen. All the had been made clear about all the facts of this system and a negotiation form has been signed by all

showing their consent. Each subject performs ten trials for each movement in total they perform 40 trials for all the four movements. Total ten subjects have been selected hence in total 400 trials has been performed by all the individuals. All the subjects are healthy individuals without any ailment (Figs. 11, 12 and 13).

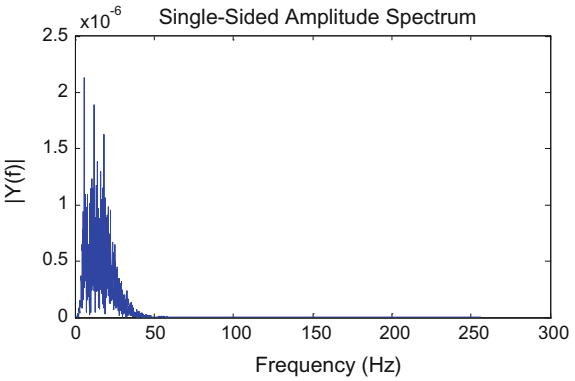


Fig. 11 Backward movement

Fig. 12 EEG based wheelchair prototype

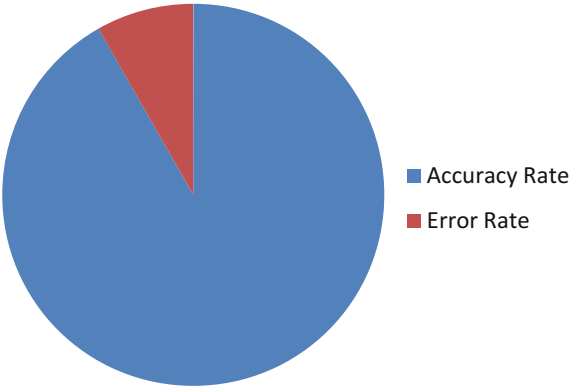


Fig. 13 Accuracy rate versus error rate

This table indicates the average accuracy of the BCI system is 91.75% with an error rate of 8.25%. A pie chart below shows the comparison of Accuracy Rate versus Error Rate (Tables 5, 6 and 7).

Table 5 Session record is shown in the table

Number of subjects	10
Trials performed by each subject for forward movement	10
Trials performed by each subject for backward movement	10
Trials performed by each subject for left movement	10
Trials performed by each subject for right movement	10
Total number of trials performed by each subject	40
Total number of trials performed by all the subjects in all the directions	400

Table 6 Table shows accuracy rate of the subjects for each type of movement

Accuracy	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Forward movement										
Correct detections (10)	9	8	8	10	8	7	10	9	9	8
Accuracy (%)	90	80	80	100	80	70	100	90	90	80
Backward movement										
Correct detections (10)	9	8	9	10	9	8	10	9	8	10
Accuracy (%)	90	80	90	100	90	80	100	90	80	100
Right movement										
Correct detections (10)	9	9	9	10	9	9	10	10	10	10
Accuracy (%)	90	90	90	100	90	90	100	100	100	100
Left movement										
Correct detections (10)	10	9	8	10	9	10	10	10	10	10
Accuracy (%)	100	90	80	100	90	100	100	100	100	100

Table 7 Table shows average accuracy rate of the system

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Total correct detections by each subject (Max. 40)	37	34	34	40	35	34	40	38	37	38
Accuracy rate (%)	92.5	85	85	100	87.5	85	100	95	92.5	95
Average number of correct detections by all the subjects								367/400		
Average accuracy rate (%)								91.75		

4 Conclusion

In this work, an attempt has been made to control the BCI based wheelchair motion using brain waves. This type of system helps paralyzed and disabled people to move as it does not require any muscular power to operate the system. In this system, stimulation of brain is done with the help of four flickering boxes, each flicker at a unique frequency, presented on the screen. The SSVEP responses generated by four different flickering frequencies which are 6, 7.5, 12, and 14 Hz are used to control a wheelchair prototype in four different directions. EEG brain signals are captured with the help of Nexus 10 machine from occipital region of the brain. Changes in the EEG patterns have been detected due to mental tasks. Hardware implementation using microcontroller board and motor drivers has been done successfully. With the implementation of above techniques, successful wheelchair control has been achieved. This system has been validated by ten subjects with multiple attempts. All the subjects participated in the experiment has shown a better accuracy rate. This proves the reliability and safety of wheelchair control. The average accuracy for forward movement is of 86%, for backward movement is 90%, for right movement is 95% and for left movement is 96%. The total values of classification accuracy vary in the range from 85–100%. Mean accuracy rate is 91.75% which indicates higher performance has been achieved. Hence SSVEP BCI is a promising way to develop a wheelchair control for disabled persons.

Declaration Authors have obtained all ethical approvals from appropriate ethical committee and approval from the subjects involved in the study.

References

1. Brainmaster Technologies: (n.d.). www.brainmaster.com/help/install.html. Accessed 19 May 2016.
2. Graimann, B., Allison, B., Pfurtscheller, G.: Brain–computer interfaces revolutionizing human–computer interaction. In: The Frontiers Collection. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-02091-9
3. Jing, W., Guanghua, X., Jun, X., Feng, Z., Lili, L., Chengcheng, H., Yeping, L., Jingjing, S.: Some highlights on EEG-based Brain Comput. Interface. http://www.paper.edu.cn
4. Yin, E., Zhou, Z., Jiang, J., Yu, Y., Hu, D.: A dynamically optimized SSVEP brain-computer interface (BCI) speller. IEEE Trans. Biomed. Eng. **67**, 1–10 (2013)
5. Waytowich, N.R., Krusienski, D.J.: Multiclass steady-state visual evoked potential frequency evaluation using chirp-modulated stimuli. IEEE Trans. Hum. Mach. Syst. **7**, 1–8 (2015)
6. Fan, X., Bi, L., Teng, T., Liu, Y.: A brain-computer interface-based vehicle destination selection system using P300 and SSVEP signals. IEEE Trans. Intell. Transp. Syst. **16**, 274–283 (2015)
7. Li, Y., Pan, J., Wang, F., Yu, Z.: A hybrid BCI system combining P300 and SSVEP and its application to wheelchair control. IEEE Trans. Biomed. Eng. **60**, 3156–3166 (2013)
8. Shyu, K.K., Lee, P.L., Lee, M.H., Lin, M.H., Lai, R.J., Chiu, Y.: Development of a low-cost FPGA-Based SSVEP BCI multimedia control system. IEEE Trans. Biomed. Circuits Syst. **4**, 125–132 (2010)

9. Wu, Z., Lai, Y., Xia, Y., Wu, D., Yao, D.: Stimulator selection in SSVEP based BCI. Med. Eng. Phys. 1079–1088 (2008). Elsevier, ChengDu, China
10. Bakardjian, H., Tanakaa, T., Cichocki, A.: Optimization of SSVEP brain responses with application to eight-command brain–computer interface. Neurosci. Lett. 34–38 (2010). Elsevier, Tokyo, Japan
11. Chang, H.C., Lee, P.L., Lo, M.T., Lee, I.H., Yeh, T.K., Chang, C.Y.: Independence of amplitude-frequency and phase calibrations in an SSVEP-based BCI using stepping delay flickering sequences. IEEE Trans. Neural Syst. Rehabil. Eng. **20**, 305–312 (2012)
12. Trejo, L.J., Rosipal, R., Matthews, B.: Brain-computer interfaces for 1-D and 2-D cursor control: designs using volitional control of the EEG spectrum or steady-state visual evoked potentials. IEEE Trans. Neural Syst. Rehabil. Eng. **14**, 225–229 (2006)
13. Hsu, H.T., Lee, I.H., Tsai, H.T., Chang, Shyu, K.K., Hsu, C.C., Chang, H.H., Yeh, T.K., Chang, C.Y., Lee, P.L.: Evaluate the Feasibility of using frontal SSVEP to implement an SSVEP-based BCI in Young, Elderly and ALS groups. IEEE Trans. Biomed. Eng. **58**, 11–23 (2015)
14. Bi, L., Fan, X., Jie, K., Teng, T., Ding, H., Liu, Y.: Using a head-up display-based steady-state visually evoked potential brain-computer interface to control a simulated vehicle. IEEE Trans. Intell. Transp. Syst. **15**, 959–966 (2014)
15. Jia, C., Gao, X., Hong, B., Gao, S.: Frequency and phase mixed coding in SSVEP-based brain-computer interface. IEEE Trans. Biomed. Eng. **58**, 200–206 (2011)
16. Noreika, A.: Controlling automated home in an eye-blink (2014). www.technology.org/2014/12/30/controlling-automated-home-eye-blink/. Accessed 12 Apr 2016
17. Pfurtscheller, G., Escalante, T.S., Ortner, R., Linortner, P., Mullerputz, G.R.: Self-Paced Operation of an SSVEP-Based Orthosis With and Without an Imagery-Based "Brain Switch:" A Feasibility Study Towards a Hybrid BCI. IEEE Trans. Neural Syst. Rehabil. Eng. **18**, 409–414 (2010)

E-Governance an Ease or Difficult to Chase in Lucknow, Uttar Pradesh

Guncha Hashmi, Pooja Khanna and Puneet Sharma

Abstract E-Governance is an appreciated initiative worldwide and also in India. The objective of E-Governance is to bridge the communication gap between government and citizens. In India, certain states are running successful E-Governance projects while in few states it is still facing challenges like lack of IT literacy, internet accessibility, etc. Before implementation of E-Governance project, a feasibility study is must to ensure probable difficulties that can hinder the success of E-Governance. This work presents a feasibility study for acceptance level of E-Governance in Lucknow the capital city of Uttar Pradesh. Research objective is to check ease or difficulty in acceptance of E-Governance in Lucknow, Uttar Pradesh. The major finding of the study says that in Lucknow majority of respondent were ready to accept E-Governance but they were unaware about such E-Governance projects running. Therefore, this feasibility study suggests a comprehensive and a large investigation to check further measures to improve the acceptance level of E-Governance.

Keywords E-Governance · G2C · Pilot survey · Feasibility study

1 Introduction

The objective of E-Governance is to provide transparent and user-friendly government. E-Governance aims to bridge the gap between services provided by Government and its stake holders. The major stakeholders of E-Governance are the

G. Hashmi (✉) · P. Khanna · P. Sharma
Department of Computer Science & Engineering, Amity University, Lucknow, India
e-mail: anshia.gunzz@gmail.com

P. Khanna
e-mail: pkhanna@lko.amity.edu

P. Sharma
e-mail: psharma9@lko.amity.edu

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_68

citizens of the country. The expectation from the government to change itself instantly into new Hi-Tech organization simply by adding "E" in it is not feasible. It is actually an ongoing exercise that needs to evaluate new framework, trained people accordingly which include common citizens and people involved in government bodies. India's population which is considered to be an asset also offers many challenges which hinder the successful implementation of any E-Governance project because net connectivity and awareness about such project cannot reach to every corner of a country along with various other issues. Apart from population, there are many other challenges that E-Governance is facing in India and one of them is low literacy rate especially IT literacy. Information Communication and Technology (ICT) is a major pillar for the establishment of any 'E' project. Any ICT based project requires strong technical and telecommunication infrastructure. India is a country with the major population living below poverty line and 60% population living in rural areas. It is a fact that problems faced by different regions are entirely different because of environment and surroundings. Hence before the launch of any E-Governance project, it is required that a proper study must be conducted to identify major challenges that could hinder the success of any E-Governance project in that particular region.

2 Background Study

Uttar Pradesh is highly populated state of our country also the fourth largest state by the economy with GDP US$120 billion [1], but along with it state also has low literacy rate which is only 57.4% according to census 2001 [2]. Uttar Pradesh took initiative for spreading E-Governance through capacity building in 2006 to provides effective delivery of its offered services and transparency to their citizens [2].

UP government got their website in almost every department so that citizen can access information by themselves without any middleman. By the help of E-Governance exchange of information with the citizens, government parties became easier and more efficient. E-Governance services in villages of all district in Uttar Pradesh is been delivered by CSC (Common Services Center) these centers are internet enabled in order to serve the common public [1].

E-Government Major Components

I. G2C (Government to Citizen) component of E-Government involves interaction between government and individuals

II. G2B (Government to Business) component of E-Government involves interaction between business entities and government.

III. G2G (Government to Government) component of E-Government involves interaction among officials within government bodies [3].

Here authors are much concerned about G2C services and how to improve its accessibility. There are some projects in Uttar Pradesh that are running successfully.

Lokhvani
E-Suvidha
E-Seva
Koshvani
Jan Suvidha Kendra
Shrishti
Bhulekh [4].

Lokhvani was initiated in November 2004 at district Sitapur, Uttar Pradesh. Objective of this project was to provide single window for providing solutions to common people like land record maintenance, handling of grievances, etc. Major part (88%) of this district is village also their literacy rate were 38% only. Therefore, it was kept in mind while designing of the website; hence chosen language was the local Hindi language so that people could find ease in accessing websites [5]. However, in 1988 central government drafted a National IT policy for the improved computer connectivity at village level and it was recommended to create infrastructure facilities to improve data communication [6].

In the recent, past many of survey and study has been conducted in order to check the acceptance and challenges facing by E-Governance in several parts of the India. In one of the study, it is found that there exist a correlation between E-Governance and reduction in corruption. The survey was conducted in Ethiopia and Fiji consisted of 800 participants who supported the hypothesis that E-Governance helps in smoothing the relationship between common citizens and government bodies also help in reducing corruption by eliminating middlemen [7]. Adarsh et al studied E-Governance facilities provided by Gujarat government and found awareness about websites is minimal [8]. Soubam et el studied G2C information services under E-Governance initiative and conducted pilot survey in Manipur in order to analyze the usability of the services by the citizens and found most of the participants were not familiar with E-Governance offered service also the accessibility is found to be poor and level of awareness among different age group for the G2C information services was relatively low [9]. Guna et al in one of the study conducted upon the effectiveness of E-Governance in Odisha state found that awareness about E-Governance services among urban areas is high as compared to the rural area also it is found that people there are giving priorities to avail E-Governance services as compared to manual services [10].

For the successful implementation of E-Governance availability and accessibility are the two crucial requirements [11]. Authors in their study have taken these two into account along with acceptance and awareness to check and analyze E-Governance running condition in Lucknow, Uttar Pradesh.

3 Research Objective

The study sought to achieve the following objectives:

I. Availability of infrastructure in Lucknow, Uttar Pradesh, like electronic gadgets, internet connection, internet speed availability, and electricity status to support successful implementation and working of E-Governance.
II. Accessibility of E-Governance websites, like how much ease in accessing websites, GUI, and language of websites.
III. Acceptance of E-Governance in terms of trust factor while giving personal details online, awareness of the keywords used in the website, also how easily individual adopt themselves with the latest upcoming technology.
IV. Awareness program about E-Governance projects in Lucknow, Uttar Pradesh, how frequently it is organized and attended by citizens also what should be the suitable mode of awareness program in Lucknow, Uttar Pradesh.

4 Research Methodology

The descriptive survey research design was adopted for the study [12]. The questionnaire methodology is used for data collection. The population consisted of mixed citizens of Lucknow, Uttar Pradesh on the basis of age, gender, qualification, occupation, monthly income, also from different regions of Lucknow, Uttar Pradesh. The sample consisted of 200 participants. The researchers constructed a questionnaire entitled E-Governance "An Ease or Difficult to Chase" which was used for data collection. The instrument is made up of four sections. Section A sought the basic details of the participants; section B contains questions requiring participants to rate the availability of infrastructure in terms of electronic gadgets, internet connectivity, internet reach, means of internet service they are using and also the purpose of using it, section C contains questions regarding accessibility of E-Governance websites, number of time visited and comparison in ease of accessing E-Governance with E-Commerce websites, while section D contains questions requiring participants to rate whether they know any awareness program for E-Governance, their visit to such programs and from whom they got to know about latest E-Center launch in Lucknow, Uttar Pradesh. While analysis of data it was found that 50% male and 50% female participants were there with 50% of Private Employee by occupation and most of the participants were Under Graduate (36%) by qualification.

The instrument was validated by expert opinion and the reliability of the instrument was ensured through a pilot study, which was conducted within an interval of 4 weeks. Data obtained from the study were analyzed using MS-Excel.

5 Data Presentation and Analysis

5.1 Availability of Infrastructure

As we are aware of well-known fact that success of E-Governance is dependent on, availability of technical infrastructure and telecommunication infrastructure, therefore, it is important to check the availability of technical and telecommunication infrastructure in Lucknow, Uttar Pradesh. Technical infrastructure includes the availability of smart devices and telecommunication infrastructure is the connectivity with the internet and the internet coverage range. Therefore in order to access E-Governance website one is required to own their smart devices which include Personnel Computers, Laptops, Mobile Phone, Tablet etc. along with internet connectivity by any mode like WIFI, broadband connection, mobile data services, net setter, etc. As such arrangements are quite expensive and expecting every citizen to afford them is impractical. To make E-Governance a success so that majority of people can utilize the benefits of E-Governance there are many working e-Centers provided by the government. Participants of this pilot study have been asked about the availability of both technical and telecommunication infrastructure in Lucknow. Table 1 represents major percentages of the respondent who have smart devices and internet connection. However in Lucknow availability of network for internet connection is a major issue. Only 31% of the respondent has no issue with network connectivity, rest of them mostly faces problem in availability of proper internet connection in Lucknow.

5.2 Acceptance of E-Governance

During the survey, the participants in Lucknow were asked about how many times they have visited any E-Governance websites, especially like UPPCL, BHULEKH, LOKHVANI. Table 2 represents the percentage of respondents who never visited any E-Governance sites. Respondents were categorized age wise. For study we categorized age into four groups: 18–25, 26–40, 41–55, and 56 above. It is very clear from the Table 2 that a major percentage of respondents of group age 41–55 and above 56 never visited the E-Governance websites. Literacy has a major impact over the acceptance level of E- Governance in Lucknow and it is clearly visible from the data of Table 3. Table 3 shows that percentage of respondents who never

Table 1 Availability of infrastructure in Lucknow, Uttar Pradesh

Availability	Percentage (%)
Smart devices	83
Internet connection	63
Network	31

Table 2 No of participants never visited E-Governance website on the basis of age groups

Age	Total participants	Never visited E-Governance website
18–25	51	18(35%)
25–40	63	20(32%)
40–55	47	33(70%)
55–Above	37	27(69%)
Total	200	97(48.5%)

Table 3 No of participants never visited E-Governance website on the basis of category qualification

Qualification	Total participants	Never visited E-Governance website
Never went school	8	8(100%)
Till 5th	5	5(100%)
Till 10th	9	8(89%)
Till 12th	33	32(97%)
Till U.G	70	28(40%)
Till P.G	58	15(26%)
Above P.G	17	2(12%)
Total	200	98(49%)

visited E-Governance website is very high in case of participants with less qualification (never went school, till 5th standard to till 12th standard). Participants with higher qualification are more frequent with E- Governance websites. Table 4 represents occupation-based classification of respondent who never visited the E-Governance websites. Table 4 clearly shows that a high percentage of the respondent in all categories who never visited the E-Governance website except private employee respondent who are accessing E-Governance site majorly(72%) in Lucknow.

Table 4 No of participants never visited E-Governance website on the basis of category occupation

Occupation	Total participants	Never visited E-Governance website
Student	23	14(61%)
Self employed	10	7(70%)
Pvt. employee	101	28(28%)
Govt. employee	13	7(54%)
Businessman	12	4(37%)
Unemployed	41	38(93%)
Total	200	98(49%)

Table 5 No of participants never visited E-Governance website on the basis of category gender

Gender	Total participants	Never visited E-Governance website
Male	100	35(35%)
Female	100	63(63%)
Total	200	98(49%)

Gender has also major impact over acceptance level as out of 100 female participants 63% never visited the E-Governance websites. Data presented in Table 5 clearly specifies that in Lucknow male respondents are more frequent with use of E-Governance websites than female respondents.

5.3 Accessibility of E-Governance

Although Indian E-Governance websites are bound to WACG 2.0 guidelines, still authors asked respondent, regarding difficulties they faced while accessing E-governance sites. They have been questioned on their understanding of the language used to present the contents of websites and requirements of technical help in order to access E-Governance website. 31% of the respondents found the language used for the website is difficult during accessing the contents of websites. Also, 22% of respondent also require a need of a technical guide to surfing the E-Government sites. When respondent were asked to compare E-Governance sites and other available E-Commerce sites 16% of found E-Commerce has better GUI than E-Government sites.

5.4 Awareness Program for E-Governance

Participants during this survey had a strong opinion for the awareness of latest E-Governance services. Respondent were asked to mention whether they know any such program for awareness about E-Governance projects in which 68% of people answered no, whereas 12% have no idea about any such awareness program running by the government and only 20% of them accepted that they know such programs conducting in Lucknow, Uttar Pradesh. In answer to a question that how many times they visited to such awareness program conducted, 82% people admitted they never visited such programs whereas 14% of participants mentioned they have visited their sometime. They were also asked to choose the best mode for the awareness programs for E-Governance projects in which most of them replied social media, television, and radio would be preferred mode for such awareness.

6 Findings of Survey

As our study suggests infrastructure is not a challenge for citizens of Lucknow, Uttar Pradesh. Also, the internet connectivity is better here, therefore, it is to find what are the actual reason why people are not using or accessing E-Governance instead they are using internet for other purpose like chatting, social media, net banking, results, downloading, etc.

I. From the survey, it is found that 82% of female are having smart devices with net connectivity among them 63% of female visited none of the E-Governance websites instead they are using it for other purpose, like chatting, social media etc.
II. Authors also found that student who participated in the survey have 100% availability of infrastructure but their awareness regarding E-Governance is very low also 78% of students have no idea about any awareness program running in Lucknow, Uttar Pradesh.
III. It is analyzed by the data that 54% people who are postgraduate by qualification find accessing E-Governance sites as a web application easier as compared to using it as mobile application.

7 Suggestions

Findings of this survey show participants are experiencing certain issues and challenges with E-Governance. By keeping those in mind authors suggested some key areas to improve upon which includes:

I. The government of Uttar Pradesh may conduct awareness program about the latest upcoming E-Governance projects so that citizens can be fully aware of its use and advantages. Preferable modes for awareness according to participants are Social Media, Hoardings, Newspaper or Radio.
II. As there are many IT colleges and universities present in Lucknow, Uttar Pradesh, therefore, they may take initiative to educate people about how to access E-Governance website and they might aim to make at least one person among family as IT literate which could solve the problem of middleman between common citizens and government bodies. In such colleges or universities, infrastructure and human resource will not be a problem.
III. Government should keep track and monitor the regular transaction, also there should be two way communication between government and citizens on their opinions on policies, performance of government through feedbacks so that citizens can feel the transparency and can build a faith in government bodies knowing that there are people other side who are actually listening to their problem.

IV. Detailed study with more number of participants will be more helpful and could clearly state the problem why acceptance is not much for E-Governance in spite of availability of infrastructure in Lucknow, Uttar Pradesh

8 Conclusion

On the basis of the feasibility study conducted the authors suggest that for an effective adoption of E-Governance websites a more comprehensive study is required to be conducted to find out probable reasons of lower acceptance level of E-Governance. The study says that IT infrastructure provided in the city is good and most of the citizens are using internet facility for other utilities. Authors suggest that a more comprehensive survey needs to be conducted to analyze the reasons of lower acceptance in Lucknow, Uttar Pradesh, and how to increase awareness among the citizens.

References

1. Singh, A.: E-governance—Initiatives in Uttar Pradesh. Asian J. Technol. Manage. Res. **04**(01) (2014)
2. Nandan, S.: Lesson from E-government Initiatives in Uttar Pradesh
3. Shaikh, M.: E-governance in Bangladesh-survey. In: Analysis and Proposed Recommendations. 19th Telecommunications forum TELFOR (2011)
4. Saxena, S., Agarwal, D.: A review of barriers found in E-governance projects in indian states. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **5**(11) (2015)
5. Sitapur Official Website. http://sitapur.nic.in/projects_implemented.htm
6. Rao, R.: ICT and E-governance for Rural Development
7. Ndou, V.: E-government for developing countries: opportunity and challenges. Electron. J. Inf. Syst. Dev. Ctries. **18**(1), 1–24 (2004)
8. Patel, A., Patel, M., Biju, S.: A Survey on E-Governance Facility Provided by Gujarat Government. Int. J. Sci. Res. Publ. **3**(8) (2013)
9. Sophiarani, S., Singh, C.: Use of G2C information services under the E-governance initiatives in NE India: a pilot survey. Ann. Libr. Inf. Stud. **60** (2013)
10. Netheti, G., Shrivastava, A., Shukla, R.: Analysis of E-governance services for effective communication to citizens in Odisha state. Int. Res. J. Eng. Technol. **02**(02) (2015)
11. Abanumy, A., Badi,A., Mayhew, P.: E-government website accessibility: in-depth evaluation of Saudi Arabia and Oman. Electron. J. E-Gov. **3**(3) (2005)
12. Malhotra, N.K.: Marketing research: An applied orientation, 5/e. Pearson Education India (2008)

Domain-Based Search Engine Evaluation

Nidhi Bajpai and Deepak Arora

Abstract The user can access web search engine at any time from anywhere to find information present on the internet or the World Wide Web. The internet is full of millions of pages which provide information related to user's query. The criticality lies in the fact that results of search engines should be a set of pages or links which are reliable and satisfies the need of the user. The efficiency of a search engine to understand users' query and provide results also forms the basis of evaluation. The method adopted in this paper establishes Efficiency, Satisfaction, and Reliability as three important parameters to evaluate search engine. Authors also introduce and define Trust Factor and Degree of Verification in this research. The aim of this study is to determine the efficiency of Search Engines, user satisfaction and reliability of search engine results based on an experiment which was conducted on working professionals employed in various domains like software companies, law firms, banks, educational institutes, Government, etc.

Keywords Search engine · Evaluation · Efficiency · Satisfaction
Reliability

1 Introduction

The typical process of web crawling, indexing and searching [1] occurs whenever a user enters a search query. The exact methods of ranking of search engine result pages differ from search engine to search engine. From users' perspective, the importance of search engine usage is related to user's satisfaction and relevance of

N. Bajpai (✉) · D. Arora
Department of Computer Science & Engineering, Amity School of Engineering,
Amity University, Lucknow, India
e-mail: nidhibajpai07@gmail.com

D. Arora
e-mail: deepakarorainbox@gmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1_69

search results. Based on these perspectives, the user decides "which" search engine to use for optimal satisfaction and relevant results.

In order to make search engines more efficient and valuable to the users, the authors proposed the study of search engine evaluation. The authors determined and estimated three parameters for the search engine evaluation. The parameters for search engine evaluation are search engine efficiency, user satisfaction and reliability of search engine results. In order to evaluate search engine, it is important to understand user behavior and perspective hence this study is based on user inputs. The authors have presented parameters for evaluating search engines. These parameters are studied and analyzed and their domain wise analysis is performed. Different domains include software companies, law firms, banks, education, government offices and others.

2 Identification of Parameters of Evaluation and Experimental Setup

In order to derive these preferences, the authors conducted an online experiment on working professionals employed in different domains as indicated above. In this experiment, few questions related to search engine were presented to the participants. Based on the input received from the participants, the authors have determined different parameters based on which domain-based search engine evaluation is performed. These different parameters for search engine evaluation are discussed in the headings below.

2.1 Efficiency of Search Engine

The efficiency of a retrieval system can be evaluated based on how much aid it provides to its users to perform their task effectively, hence in order to determine the efficiency of the search engine, the perspective of users rating on search engine results must be taken into account [2]. According to authors, In order to determine the efficiency of search engine, the users need to be analyzed for two aspects:

- Efficiency of search engines to give accurate results.
- Efficiency of search engines to understand query of the user.

In order to determine the first aspect of efficiency, the authors analyzed the users regarding "If their preferred search engine is able to give them accurate results". The findings to this aspect are shown in Table 1 which states that according to 15% of users their search engine is 100% efficient, i.e., it is able to give accurate results always. According to 80% of users, their search engine is able to give them accurate results most of the time while only 5% of users said that their search engine is able

to give accurate results only half of the time. Nobody opted for less than half of the time or never an option.

Similarly, for determining the second aspect of search engine efficiency, the authors analyzed the users regarding "If their preferred search engine is able to understand their query accurately." The findings to this aspect as shown in Table 2 shows that according to 91.7% of users their search engine is able to understand their query accurately while 8.3% of users feel that their search engine is not able to understand their query accurately.

The above-discussed aspects of search engine show that majority of users rate their preferred search engine to be highly efficient.

2.2 User Satisfaction

One of the prime objectives of search engines is to satisfy their users with efficient search engine results. User satisfaction is one of the important factors for evaluating search engines. Satisfaction is a complex parameter and depends on various factors [3]. The authors define user satisfaction as a measure of how search results and information provided by the search engine meet or exceed user expectations.

Satisfaction is a user-dependent parameter hence to determine user satisfaction, the authors analyzed the users regarding how much satisfied they feel by the search engine results provided by their search engines. The findings to this aspect as shown in Table 3 shows that 46.7% of users rate their satisfaction level to be more than 95%. 47.5% of users rate their satisfaction level to be more than 75% and less than 95% while only 5.8% of users rate their satisfaction level between 50 and 75%. Nobody rated the user satisfaction level lesser than 50%.

The above analysis regarding user satisfaction states that 94.2% of users rate their satisfaction level as more than 75%. This indeed establishes that working professionals in different domains are quite satisfied with the search engine results.

Table 1 Findings for "is your search engine able to give you accurate results?"

	Frequency	Percent	Cumulative percent	Variable encoding
Never	0	0	0	0
Less than half of time	0	0	0	1
Half of the time	6	5.0	5.0	2
Most of the time	96	80.0	85.0	4
Always	18	15.0	100.0	5
Total	120	100.0		

Table 2 Is your search engine able to understand your query accurately?

	Frequency	Percent	Valid percent	Cumulative percent
No	10	8.3	8.3	8.3
Yes	110	91.7	91.7	100.0
Total	120	100.0	100.0	

Table 3 Findings for user satisfaction

	Frequency	Percent	Valid percent	Cumulative percent	Variable encoding
Less than 25%	0	0	0	0	1
Between 25 and 50%	0	0	0	0	2
More than 50%	7	5.8	5.8	5.8	3
More than 75%	57	47.5	47.5	53.3	5
More than 95%	56	46.7	46.7	100.0	6
Total	120	100.0	100.0		

2.3 Reliability of Search Engine Results

The Internet is full of information but the authenticity and reliability of source from which information is coming are necessary. Search results that are derived from search engines are not always reliable. In order to check the quality of website content and produce reliable search results, search engine needs to put a lot of cost and effort for the qualitative analysis [4]. Along with this qualitative analysis, it also requires human-like reasoning to provide reliable search results. Till date, we do not have search engines which provide complete reliability to the users.

In order to analyze users' perspective on the reliability of search engines, authors defined and analyzed two aspects of reliability.

- Trust Factor
- Degree of Verification

Authors define Trust Factor as the measure of users' trust in search engine results. In order to analyze this factor, authors analyzed users regarding "Do the users trust the results of search engine completely?" The findings to Trust factor as shown in Table 4. States that 57.5% of users trust the results of search engine completely while 42.5% of users do not trust the results of search engine completely.

Authors define Degree of Verification as a number of times for which a user verifies the search engine results from authentic source of information. In order to determine the degree of Verification, the authors analyzed users regarding "Do the users verify search engine results from other sources? "The findings to it as shown in Table 5. States that Only 4.2% of users always verify the search engine results

Table 4 Findings for "do the users trust the results of search engine completely"

	Frequency	Percent	Valid percent	Cumulative percent	
No	51	42.5	42.5	42.5	1
Yes	69	57.5	57.5	100.0	2
Total	120	100.0	100.0		

Table 5 Findings for "do the users verify search engine results from other sources"

	Frequency	Percent	Valid percent	Cumulative percent	Variable encoding
Always	5	4.2	4.2	4.2	1
Never	19	15.8	15.8	20.0	2
Not always but yes	96	80.0	80.0	100.0	3
Total	120	100.0	100.0		

from other sources while 15.8% of users never verify the search engine results from other sources. Rest of the 80% of users verify the search engine results if required. This shows that degree of verification is quite low.

3 Methodology Used

3.1 Stage I—Qualitative Research

In this study, firstly the authors have used qualitative research methodology by doing a detailed study about search engines. The exploratory research was done aimed to achieve a better understanding of the topic. Literature Review technique was used to gain a better understanding of the topic which helped further to frame research questions. This exploratory research step will aid in the more powerful analysis of the subject at later stages. This step is important to have a better understanding of the topic, users' perspectives and future implications of the topic and provides better insight into the topic [5].

3.2 Stage II—Quantitative Research

A Quantitative research was followed which used the online survey methodology. A questionnaire was designed using Google forms and circulated online to working professionals in different domains like Software Company, Government, Banking, Legal firms, Education, etc. The responses collected from the working professionals were analyzed using the software GNU PSPP software [5].

4 Results and Discussions

This research work is based on an experiment conducted online amongst working professionals. The sample consisted of 120 working professionals from different domains which include 55 professionals working in different software companies, 13 law professionals practicing independently or in legal firms, 13 professionals working in Government services, 12 professionals working in the education field, 11 professionals working in different Banks and 16 professionals in 'other' domains. The sample includes 76 male professionals and 44 female professionals [6–12].

In order to evaluate search engines, authors evaluated the efficiency of search engines. According to this research based on user inputs and variable encoding done as shown in Table 1, the average value of Efficiency of Search Engine is 4.05 (Table 6) which means that search engines are able to give accurate results most of the time.

Domain wise analysis of the efficiency of search engines is shown in Fig. 1. Which states that mean value of efficiency of search engines is maximum in Banking domain, i.e., 4.1818. The mean value of efficiency is minimum in Legal firms, i.e., 3.8462. These mean values are calculated based on Variable encoding as shown in Table 1.

The mean value of efficiency of search engine for Male users is 4.0263 and for female users is 4.0909. Hence, there is no considerable difference for efficiency of search engines as rated by Male users and Female users.

The analysis of efficiency of search engine based on Experience level of professionals working in different domains is in shown in Fig. 2. Which states that users of medium level experience or users belonging to top-level positions in organizations have a higher efficiency rating of search engines as compared to novice or fresher's. The mean value of efficiency as rated by Novice users is 3.55.

Domain wise analysis of mean value of user satisfaction is shown in Fig. 3. Which states that in all the domains users' satisfaction level is more than 5, i.e., more than 75%. Users in Government domain have maximum satisfaction level, i.e., 5.4615 while users in Legal firms have a minimum level of satisfaction level, i.e., 5.0769.

Table 6 Table showing descriptive statistics for efficiency of search engine

	N	Minimum	Maximum	Mean	Std. deviation	Variance
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic
Efficiency	120	2.00	5.00	4.0500	0.59196	0.350
Valid N (list wise)	120					

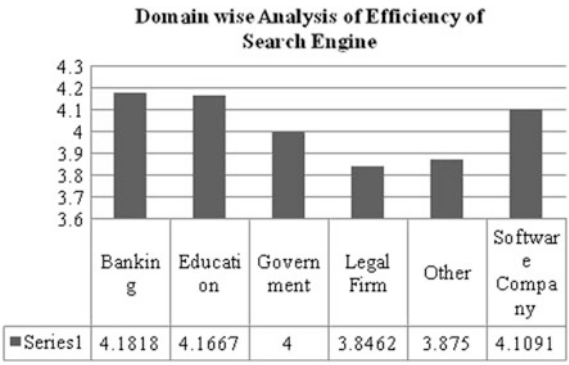


Fig. 1 Domain wise analysis of efficiency of the search engine

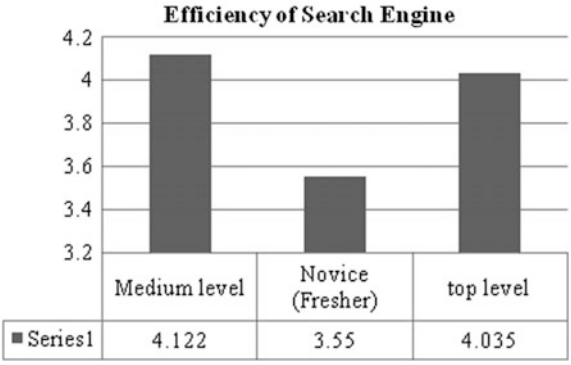


Fig. 2 Analysis of efficiency based on experience level

In order to determine reliability, as discussed above, authors define Trust Factor and Degree of Verification. The findings to Trust Factor as shown in Fig. 4. States that value of the mean value of Trust Factor is maximum for Education domain, i.e., 1.83 and minimum for Government domain, i.e., 1.31. The mean values of a trust factor for Banking, Legal firm and Software companies are 1.36, 1.38 and 1.65 respectively. The mean values are calculated based on variable encoding as shown in Table 4.

The findings to the degree of verification as shown in Fig. 5. States that only 4.2% of users always verify the search engine results from other sources while 15.8% of users never verify the search engine results from other sources. Majority of users, i.e., 80% of users verify the search engine results if required. This states that degree of verification is quite low.

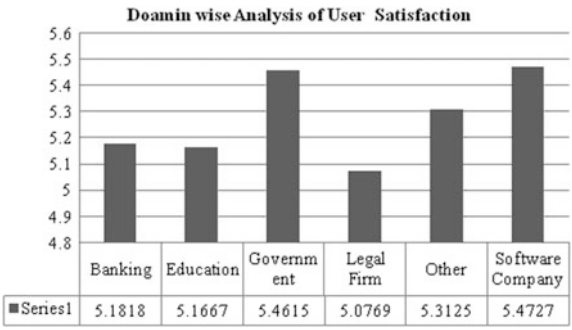


Fig. 3 Domain wise analysis of user satisfaction

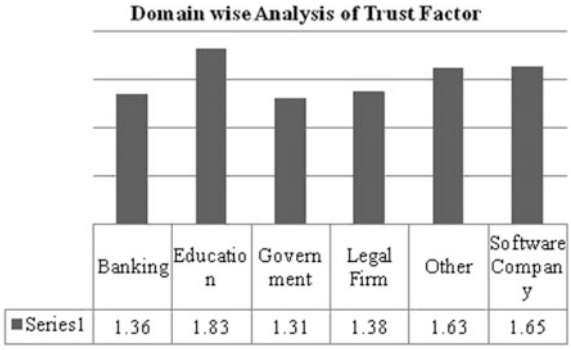


Fig. 4 Domain wise analysis of trust factor

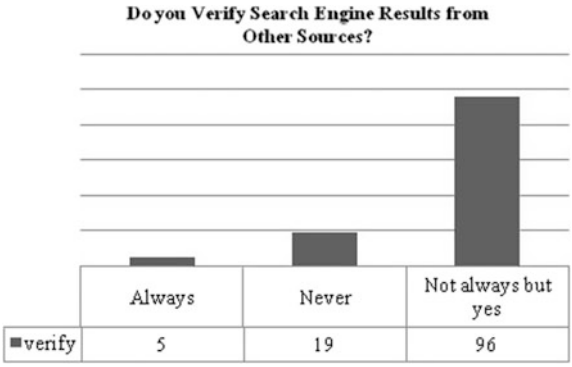


Fig. 5 Results for degree of verification

5 Conclusion

In this research work, authors have determined three parameters for search engine evaluation. These three parameters are the efficiency of the search engine, reliability of search engine results and user satisfaction. As shown in results, 95% of users say that their search engine is able to give accurate results most of the time. Results also

show that 91.7% of users agree that their search engine is able to understand their query accurately. Thus, that majority of users rate their preferred search engine to be highly efficient. Results regarding users' satisfaction states that that 94.2% of users rate their satisfaction level as more than 75%. An almost irrefutable conclusion of the author's evaluation search engine states that search engines are highly efficient and users' satisfaction level is quite high.

The mean value of efficiency of search engines is maximum in Banking domain, i.e., 4.1818 while its minimum for Legal firms, i.e., 3.8462. There is no considerable difference for the mean value of efficiency of search engines between male users and female users. Users in Government domain have maximum satisfaction level, i.e., mean value of 5.4615 while users in Legal firms have a minimum level of satisfaction level, i.e., mean value of 5.0769.

Authors also defined Trust Factor and Degree of verification to determine Reliability of search engine results. As shown in results, Trust factor is approximately half and degree of verification is quite low. This states that Reliability of search engine results is still a critical issue for search engine evaluation. Mean value of Trust Factor is maximum for Education domain, i.e., 1.83 and minimum for Government domain, i.e., 1.31.

Authors state that these three parameters of search engine evaluation are interdependent on each other. Increased Reliability of search engine results increases users' satisfaction and vice versa. Similarly, the efficient results are directly related to increased users' satisfaction. The results of this experiment will be very helpful for the upcoming researchers toward increasing the efficiency and opening new dimensions of different search algorithm/techniques, by adding more intelligence to upcoming browsers in the market. The future scope of this study includes conducting the same experiment on larger sample size by including more different domains for further analysis and domain-specific understanding.

Declaration Authors have obtained all ethical approvals from appropriate ethical committee and approval from the subjects involved in this study.

References

1. Jawadekar, W.: Knowledge Management. Tata McGraw-Hill Education, New Delhi (2011)
2. Ingwersen, P., Järvelin, K.: The Turn Integration of Information Seeking and Retrieval in Context. Springer, Dordrecht (2005)
3. Bruce, H.: User satisfaction with information seeking on the internet. J. Am. Soc. Inf. Sci. **49**, 541–556 (1998)
4. Wai, L.: System and Method for Development of Search Success Metrics (2010)
5. Bajpai, N., Arora, D.: An estimation of user preferences for search engine results and usage. In: Advances in Intelligent Systems and Computing Series. Springer, n.d. (in-press)
6. Chu, H., Rosenthal, M.: Search engines for the world wide web: a comparative study and evaluation methodology. In: Proceedings of the 59th Annual Meeting of the American Society for Information Science (1996)
7. Hunter, J., Schmidt, F.: Methods of Meta-Analysis. Sage, Thousand Oaks, California (2004)

8. Bar-Ilan, J.: Data collection on the Web for informetric purposes: a review and analysis. Scientometrics. (2001)
9. Lewandowski, D.: New Perspectives on Web Search Engine Research. Emerald Publishing, UK (2012)
10. Cooper, W.: Expected search length: a single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. Amer. Doc. **19**, 30–41 (1968)
11. Frøkjær, E., Hertzum, M., Hornbæk, K.: Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, The Hague, Amsterdam (2000)
12. Al-Maskari, A., Sanderson, M.: A review of factors influencing user satisfaction in information retrieval. J. Am. Soc. Inf. Sci. **61**, 859–868 (2010)

Author Index

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 564,
https://doi.org/10.1007/978-981-10-6875-1