

Advances in Intelligent Systems and Computing 683

Sabu M. Thampi
Sushmita Mitra
Jayanta Mukhopadhyay
Kuan-Ching Li
Alex Pappachen James
Stefano Berretti *Editors*

Intelligent Systems Technologies and Applications

 Springer

Advances in Intelligent Systems and Computing

Volume 683

Series editor

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

About this Series

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.

Advisory Board

Chairman

Nikhil R. Pal, Indian Statistical Institute, Kolkata, India

e-mail: nikhil@isical.ac.in

Members

Rafael Bello Perez, Universidad Central “Marta Abreu” de Las Villas, Santa Clara, Cuba

e-mail: rbellop@uclv.edu.cu

Emilio S. Corchado, University of Salamanca, Salamanca, Spain

e-mail: escorchado@usal.es

Hani Hagrass, University of Essex, Colchester, UK

e-mail: hani@essex.ac.uk

László T. Kóczy, Széchenyi István University, Győr, Hungary

e-mail: koczy@sze.hu

Vladik Kreinovich, University of Texas at El Paso, El Paso, USA

e-mail: vladik@utep.edu

Chin-Teng Lin, National Chiao Tung University, Hsinchu, Taiwan

e-mail: ctlin@mail.nctu.edu.tw

Jie Lu, University of Technology, Sydney, Australia

e-mail: Jie.Lu@uts.edu.au

Patricia Melin, Tijuana Institute of Technology, Tijuana, Mexico

e-mail: epmelin@hafsamx.org

Nadia Nedjah, State University of Rio de Janeiro, Rio de Janeiro, Brazil

e-mail: nadia@eng.uerj.br

Ngoc Thanh Nguyen, Wroclaw University of Technology, Wroclaw, Poland

e-mail: Ngoc-Thanh.Nguyen@pwr.edu.pl

Jun Wang, The Chinese University of Hong Kong, Shatin, Hong Kong

e-mail: jwang@mae.cuhk.edu.hk

More information about this series at <http://www.springer.com/series/11156>

Sabu M. Thampi · Sushmita Mitra
Jayanta Mukhopadhyay · Kuan-Ching Li
Alex Pappachen James · Stefano Berretti
Editors

Intelligent Systems Technologies and Applications

 Springer

Editors

Sabu M. Thampi
School of CS/IT
Indian Institute of Information Technology
Trivandrum, Kerala
India

Sushmita Mitra
Machine Intelligence Unit
Indian Statistical Institute
Kolkata
India

Jayanta Mukhopadhyay
Department of Computer Science and
Engineering
Indian Institute of Technology
Kharagpur, West Bengal
India

Kuan-Ching Li
Xiamen University
Xiamen
China

Alex Pappachen James
Department of Electrical and Electronic
Nazarbayev University
Astana
Kazakhstan

Stefano Berretti
Dipartimento di Ingegneria
Università degli Studi di Firenze
Firenze
Italy

ISSN 2194-5357

ISSN 2194-5365 (electronic)

Advances in Intelligent Systems and Computing

ISBN 978-3-319-68384-3

ISBN 978-3-319-68385-0 (eBook)

<https://doi.org/10.1007/978-3-319-68385-0>

Library of Congress Control Number: 2017954902

© Springer International Publishing AG 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

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

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature

The registered company is Springer International Publishing AG

The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This volume of proceedings provides an opportunity for readers to engage with a selection of refereed papers that were presented at the Third International Symposium on Intelligent Systems Technologies and Applications (ISTA'17). ISTA aims to bring together researchers in related fields to explore and discuss various aspects of intelligent systems technologies and their applications. This edition was hosted by Manipal Institute of Technology, Manipal University, Manipal, India, during September 13–16, 2017. ISTA'17 was colocated with the Second International Conference on Applied Soft Computing and Communication Networks (ACN'17).

All submissions were evaluated on the basis of their significance, novelty, and technical quality. A double-blind review process was conducted to ensure that the author names and affiliations were unknown to the TPC. These proceedings contain 34 papers selected for presentation at the symposium.

We are very grateful to the many people who helped with the organization of the symposium. Our sincere thanks go to all authors for their interest in the symposium and to the members of the Program Committee for their insightful and careful reviews, all of which were prepared on a tight schedule but still received in time. The conference could not have happened without the commitment of the Local Organizing Committee, who helped in many ways to assemble and run the conference. We are grateful to the General Chairs for their support. We express our most sincere thanks to all keynote speakers who shared with us their expertise and knowledge. Finally, we would like to acknowledge Springer for active cooperation and timely production of the proceedings.

Sabu M. Thampi
Sushmita Mitra
Jayanta Mukhopadhyay
Kuan-Ching Li
Alex Pappachen James
Stefano Berretti

Program Chairs

Kuan-Ching Li	Providence University, Taiwan
Alex Pappachen James	Nazarbayev University, Kazakhstan
Ben Goertzel	Aidyia Limited, Hong Kong
Farid Nait-Abdesselam	Paris Descartes University, France

Workshop Chairs

El-Sayed El-Alfy	King Fahd University of Petroleum and Minerals, Saudi Arabia
Sebastian Basterrech	VSB-Technical University of Ostrava, Czech Republic

Steering Committee Chair

Sabu M. Thampi	IIITM-Kerala, India
----------------	---------------------

Organizing Chair

Hareesha K.S	Manipal Institute of Technology (MIT), Manipal University, India
--------------	---

Organizing Co-chairs

Ashalatha Nayak	Manipal Institute of Technology, Manipal University
Balachandra	Manipal Institute of Technology, Manipal University

Organizing Secretaries

Renuka A	Manipal Institute of Technology, Manipal University
Preetham Kumar	Manipal Institute of Technology, Manipal University
Poornima PK	Manipal Institute of Technology, Manipal University

TPC Members/Additional Reviewers

Phan Cong-Vinh	NTT University, Vietnam
Thanh D. Nguyen	Banking University of Ho Chi Minh City, Vietnam
Tri-Thanh Nguyen	Vietnam National University, Hanoi, Vietnam
Hai Pham	Hanoi University of Science and Technology, Vietnam
Mustafa Jaber	Nant Vision Inc., USA
Ankit Chaudhary	Northwest Missouri State University, USA
Ninoslav Marina	Princeton University, USA
Gang Wang	Intelligent Fusion Technology, Inc., USA
Harishchandra Dubey	The University of Texas at Dallas, USA
Mahdin Mahboob	Stony Brook University, USA
Hossein Malekinezhad	Michigan Technological University, USA
Tae (Tom) Oh	Rochester Institute of Technology, USA
Anthony Tsetse	Northern Kentucky University, USA
Huasen Wu	University of California, Davis, USA
Xia Li	Qualcomm, USA
Sai Ganesh Sitharaman	Zscaler, Inc., USA
Taha Selim Ustun	Carnegie-Mellon University, USA
Abdel-Hameed Badawy	New Mexico State University, USA
J. Mailen Kootsey	Simulation Resources, Inc., USA
Eduard Babulak	Fort Hays State University, USA
Akshaye Dhawan	Ursinus College, USA
Eun-Sung Jung	Hongik University, USA
Rajgopal Kannan	University of Southern California, USA
Prabhaker Mateti	Wright State University, USA
Mariofanna Milanova	University of Arkansas at Little Rock, USA
Dibyendu Mukherjee	Duke University, USA
Vamsi Paruchuri	University of Central Arkansas, USA
Sandeep Reddivari	University of North Florida, USA
Xiaoyuan Suo	Webster University, USA
Wei Tian	Illinois Institute of Technology, USA
Yuanzhang Xiao	Northwestern University, USA
Peng Zhang	Stony Brook University, USA
Sumer Can	Diodes, Inc., USA
Pin-Yu Chen	IBM T.J. Watson Research Center, USA
Son Doan	Kaiser Permanente, USA
Yuhuan Du	Dropbox Inc., USA
Vahid Khalilzad-Sharghi	Medical Imaging Solutions USA, USA
Vishnu Pendyala	Santa Clara University, USA
Tatsuya Suda	University Netgroup Inc., USA
Xiaochuan Wang	Intel Corp., USA

Martine Wedlake	IBM, USA
Kishore Yalamanchili	Google, USA
Guangjie Huang	Auburn University, USA
Chiranjib Sur	University of Florida, USA
Varghese Vaidyan	Iowa State University, USA
Rafael Sotelo	Universidad de Montevideo, Uruguay
Nik Bessis	Edge Hill University, UK
Erol Gelenbe	Imperial College London, UK
Cathryn Peoples	Queen Mary University of London, UK
Hoi Leong Lee	University of Oxford, UK
Biju Issac	Teesside University, Middlesbrough, UK
Ali Al-Sherbaz	The University of Northampton, UK
Abdulkadir Alkali	Sheffield Hallam University, UK
Mustansar Ghazanfar	University of Southampton, UK
Hassan Hamdoun	University of Aberdeen, UK
Bilal Khan	University of Sussex, UK
Ehab Salahat	IEEE, United Arab Emirates (UAE)
Mohammad Al-Shabi	University of Sharjah, United Arab Emirates (UAE)
Arijit Bhattacharya	University of Dubai, United Arab Emirates (UAE)
Panos Liatsis	The Petroleum Institute, United Arab Emirates (UAE)
Anton Popov	National Technical University of Ukraine “Kyiv Polytechnic Institute,” Ukraine
Rostyslav Sklyar	Independent Professional, Ukraine
Pvl Rao	Kampala International University, Uganda
Burhan Gulbahar	Ozyegin University, Turkey
Albert Guvenis	Bogazici University, Turkey
Yasin Kabalci	Nigde University, Turkey
Haldun Ozaktas	Bilkent University, Turkey
Najeh Lakhoua	ENICarthage, Tunisia
Marwan Zouinkhi	Higher Institute of Applied Science and Technology of Sousse, Tunisia
Ibrahim Missaoui	National Engineering School of Tunis, Tunisia
Youssef Said	Tunisie Telecom, Tunisia
Zhiyuan Tan	University of Twente, the Netherlands
Nattee Pinthong	Rajabhat Rajanagarindra University, Thailand
Winai Jaikla	King Mongkut’s Institute of Technology Ladkrabang, Thailand
Manasawee Kaenampornpan	Maharakham University, Thailand
Mahasak Ketcham	King Mongkut’s University of Technology North Bangkok, Thailand
Grienggrai Rajchakit	Maejo University, Thailand
Ying-Ren Chien	National I-Lan University, Taiwan

Gwo-Jiun Horng	Southern Taiwan University of Science and Technology, Taiwan
San-Nan Lee	Vanung University, Taiwan
Sheng-Shih Wang	Minghsin University of Science and Technology, Taiwan
Yue-Shan Chang	National Taipei University, Taiwan
Mu-Song Chen	Electrical Engineering, Da-Yeh University, Taiwan
Chien-Fu Cheng	Tamkang University, Taiwan
Chih-Ming Kung	Shih Chien University, Taiwan
Chia-Hung Lai	National Cheng Kung University, Taiwan
Meng-Shiuan Pan	Tamkang University, Taiwan
Kuei-Ping Shih	Tamkang University, Taiwan
Ming-Fong Tsai	Feng Chia University, Taiwan
Antonio Cimmino	Lasting Dynamics, Switzerland
Oskars Ozolins	Acreo Swedish ICT, Sweden
Emilio Jiménez Macías	University of La Rioja, Spain
Jorge Bernal Bernabé	University of Murcia, Spain
Juan Corchado	Universidad de Salamanca, Spain
Pablo Corral	Miguel Hernández University, Spain
Hector Menendez	University College London, Spain
Jose Molina	Universidad Carlos III de Madrid, Spain
Addisson Salazar	Universidad Politécnica de Valencia, Spain
Carlos Travieso	University of Las Palmas de Gran Canaria, Spain
Jose Luis Vazquez-Poletti	Universidad Complutense de Madrid, Spain
Pradeep Kumar Gupta	University of Pretoria, South Africa
Sindiso Nleya	Computer Science Department, South Africa
Hwee Pink Tan	Singapore Management University, Singapore
Chau Yuen	Singapore University of Technology and Design, Singapore
Arun Kumar	National University of Singapore, Singapore
Md. Rabiul Islam	Nanyang Technological University, Singapore
Yilun Shang	Singapore University of Technology and Design, Singapore
Sasikumaran Sreedharan	King Khalid University, Saudi Arabia
Anton Satria Prabuwono	King Abdulaziz University, Saudi Arabia
Dushantha Nalin K. Jayakody	National Research Tomsk Polytechnic University, Russia
Anton Pljonkin	Southern Federal University, Russia
Radhakrishnan Delhibabu	KBSG, ITIS, Kazan Federal University, Russia
Elena Benderskaya	Saint-Petersburg State Polytechnical University, Russia
Dan Milici	University of Suceava, Romania
Felix Albu	Valahia University of Targoviste, Romania
Mihaela Albu	Politehnica University of Bucharest, Romania

Dan Dobrea	Technical University “Gh. Asachi,” Romania
Valentina Balas	Aurel Vlaicu University of Arad, Romania
Traian Rebedea	University Politehnica of Bucharest, Romania
Lucian Vintan	University of Sibiu, Romania
Christophe Soares	University Fernando Pessoa, Portugal
Saravanan Kandasamy	INESC TEC Porto, Portugal
Pedro Silva Girão	Instituto Superior Técnico, Portugal
Rodolfo Oliveira	Nova University of Lisbon, Portugal
Luis Teixeira	Universidade Catolica Portuguesa, Portugal
Davide Carneiro	University of Minho, Portugal
Jose Delgado	Technical University of Lisbon, Portugal
Paulo Neves	Polytechnic Institute of Castelo Branco, Portugal
Manuel Silva	ISEP/IPP-School of Engineering, Polytechnic Institute of Porto, Portugal
Iouliia Skliarova	University of Aveiro, Portugal
Tomasz Neumann	Gdynia Maritime University, Poland
Waqas Bangyal	Iqra University, Islamabad, Pakistan
Mansoor Khan	COMSATS Institute of Information Technology, Pakistan
Nouman Qadeer Soomro	Mehran University of Engineering and Technology, SZAB Campus, Pakistan
Kangqi Liu	Shanghai Jiao Tong University, P.R. China
Bin Cao	Harbin Institute of Technology, P.R. China
Xiangguo Li	Henan University of Technology, P.R. China
Jiayu Chen	Wuhan University, P.R. China
Laxmisha Rai	Shandong University of Science and Technology, P.R. China
Xiao Liang	Shanghai Fuxin Intelligent Transportation Solutions Co. Ltd., P.R. China
Pingyi Fan	Tsinghua University, P.R. China
Philip Moore	Lanzhou University, P.R. China
Mithun Mukherjee	Guangdong Provincial Key Lab of Petrochemical Equipment Fault Diagnosis, P.R. China
Zhengxing Sun	Nanjing University, P.R. China
Wei Wei	Xi’an University of Technology, P.R. China
Jiping Xiong	Zhejiang Normal University, P.R. China
Bidi Ying	Zhejiang Gongshang University, P.R. China
Shigeng Zhang	Central South University, P.R. China
Qiyang Zhao	Beihang University, P.R. China
Xian Li	Southeast University, P.R. China
Jinbei Zhang	Shanghai Jiao Tong University, P.R. China
Mhamed Bakrim	University of Cadi Ayyad Marrakech, Morocco
Hajami Abdelmajid	ENSIAS/FSTS, Morocco
Malaoui Abdessamad	Sultan Moulay Slimane University of Beni Mellal, Morocco

Soufiana Mekouar	Mohammed V University Rabat, Morocco
Karla Maria Ronquillo Gonzalez	Universidad Tecnológica de Chihuahua, Mexico
Marcelo Romero	Autonomous University of the State of Mexico, Mexico
Shireen Panchoo	University of Technology, Mauritius
Rini Akmeiliawati	International Islamic University Malaysia, Malaysia
Muataz Salih	UniMap, Malaysia
Bunseng Chan	Universiti Malaysia Sabah, Malaysia
Amir Faisal	University of Malaya, Malaysia
Thinagaran Perumal	University Putra Malaysia, Malaysia
Ahmed Almurshedi	Universiti Teknologi Malaysia, Malaysia
Hong Seng Gan	Universiti Kuala Lumpur, Malaysia
Mohammad Faiz Liew Abdullah	Universiti Tun Hussein Onn Malaysia (UTHM), Malaysia
Mohd Ashraf Ahmad	Universiti Malaysia Pahang, Malaysia
Nahrul Khair Alang Md Rashid	Universiti Teknologi Malaysia, Malaysia
Azrul Nizam Alias	Universiti Teknologi MARA, Malaysia
Asrul Izam Azmi	Universiti Teknologi Malaysia, Malaysia
Kalaivani Chellappan	Universiti Kebangsaan Malaysia, Malaysia
Rudzidatul Dziauddin	Universiti Teknologi Malaysia, Malaysia
Mohd Khair Hassan	Universiti Putra Malaysia, Malaysia
Kushsairy Kadir	Universiti Kuala Lumpur British Malaysian Institute, Malaysia
Raja Kamil	Universiti Putra Malaysia, Malaysia
C.M.R. Prabhu	Multimedia University, Malaysia
Md Jan Nordin	Universiti Kebangsaan Malaysia, Malaysia
Siti Hawa Ruslan	Universiti Tun Hussein Onn Malaysia, Malaysia
Fadzilah Siraj	Universiti Utara Malaysia, Malaysia
Abd Kadir Mahamad	Universiti Tun Hussein Onn Malaysia, Malaysia
Su Fong Chien	MIMOS Berhad, Malaysia
Min Keng Tan	Universiti Malaysia Sabah, Malaysia
Veeraiyah Thangasamy	Asia Pacific University of Technology and Innovation, Malaysia
Chitti Babu B	The University of Nottingham Malaysia Campus, Malaysia
Marwan Nafea	Universiti Teknologi Malaysia (UTM), Malaysia
Ahmad Yusairi Bani Hashim	Universiti Teknikal Malaysia Melaka, Malaysia
Faieza Abdul Aziz	Universiti Putra Malaysia, Malaysia
Shazmin Aniza Abdul Shukor	Universiti Malaysia Perlis, Malaysia
Rozzeta Dolah	Universiti Teknologi Malaysia, Malaysia
Mohammad Rajib Hasan	University Utara Malaysia, Malaysia
Mohd Hanafi Ahmad Hijazi	Universiti Malaysia Sabah, Malaysia

Ku Nurul Fazira Ku Azir	Universiti Malaysia Perlis, Malaysia
Jane Labadin	Universiti Malaysia Sarawak, Malaysia
Adidah Lajis	University Kuala Lumpur, Malaysia
Massudi Mahmuddin	Universiti Utara Malaysia, Malaysia
Mat Nawi	Universiti Pendidikan Sultan Idris, Malaysia
Rosalyn R. Porle	Universiti Malaysia Sabah, Malaysia
Hua Nong Ting	Universiti Malaya, Malaysia
Noradlina Abdullah	TNB Research Sdn Bhd, Malaysia
Chong Fook	University of Malaysia in Perlis, Malaysia
M. Hassan	Universiti Teknologi Petronas, Malaysia
Lee Hung Liew	Universiti Teknologi MARA Sarawak, Malaysia
Feliksas Kuliesius	Vilnius University, Lithuania
Michel Owayjan	American University of Science and Technology, Lebanon
Zein Al Abidin Ibrahim	Lebanese University, Lebanon
Seifedine Kadry	American University of the Middle East, Kuwait
Sa'ed Abed	Kuwait University, Kuwait
Sayed Chhattan Shah	Hankuk University of Foreign Studies, South Korea, Korea
Young Bin Kown	Chung-Ang University, Korea
Dhananjay Singh	Hankuk University of Foreign Studies, Korea
Dong-hwa Kim	Hanbat National University, Korea
Atallah AL-Shatnawi	Al al-Bayt University, Jordan
Bashar Al-Shboul	The University of Jordan, Jordan
Guanghao Sun	The University of Electro-Communications, Japan
Noriko Etani	Kyoto University, Japan
Kazuo Mori	Mie University, Japan
Akihiro Fujihara	Fukui University of Technology, Japan
Yuji Iwahori	Chubu University, Japan
Kenichi Kourai	Kyushu Institute of Technology, Japan
Dramjad Mehmood	Guangdong University of Petrochemical Technology, Japan
Hirosato Seki	Osaka University, Japan
Hideyuki Takahashi	Tohoku University, Japan
Atsushi Takeda	Tohoku Gakuin University, Japan
Luis F. Abanto-Leon	Tohoku University, Japan
Federico Tramarin	National Research Council of Italy, Italy
Angelo Genovese	Università degli Studi di Milano, Italy
Paolo Crippa	Università Politecnica delle Marche, Italy
Domenico Ciuonzo	University of Naples Federico II, Italy
Manuel Roveri	Politecnico di Milano, Italy
Gennaro Boggia	Politecnico di Bari, Italy
Massimo Ficco	Second University of Naples, Italy
Domenico Grimaldi	University of Calabria, Italy

Orazio Tomarchio	University of Catania, Italy
Aniello Castiglione	Università di Salerno, Italy
Maurizio Naldi	University of Rome “Tor Vergata,” Italy
Alba Amato	Institute for High-Performance Computing and Networking (ICAR-CNR), Italy
Flora Amato	Università degli Studi di Napoli Federico II, Italy
Marco Anisetti	Università degli Studi di Milano, Italy
Massimo Cafaro	University of Salento, Italy
Riccardo Colella	University of Salento, Italy
Mario Collotta	Kore University of Enna, Italy
Eleonora D’Andrea	University of Pisa, Italy
Ruggero Donida Labati	Università degli Studi di Milano, Italy
Ugo Fiore	University of Naples Federico II, Italy
Francesco Masulli	University of Genova, Italy
Ambra Molesini	Alma Mater Studiorum-Università di Bologna, Italy
Andrea Omicini	Alma Mater Studiorum-Università di Bologna, Italy
Giovanni Pau	Kore University of Enna, Italy
Danilo Pelusi	University of Teramo, Italy
Giovanni Pilato	Italian National Research Council, Italy
Lorenzo Mossucca	Istituto Superiore Mario Boella, Italy
Ruifeng Zhang	Letterkenny Institute of Technology, Ireland
Deepak Mehta	United Technologies Research Centre, Ireland
Safanah Raafat	University of Technology Baghdad, Iraq
Majida Alasady	University of Tikrit, Iraq
Ali Hussein Hasan	Sumer University, Iraq
Imad Mohamad	University of Baghdad, Iraq
Behrooz Razeghi	Ferdowsi University of Mashhad, Iran
Reza Atani	University of Guilan, Iran
Peyman Arebi	Technical and Vocational University-Technical and Vocational College of Bushehr, Iran
Amir Hosein Jafari	Iran University of Science and Technology, Iran
Abdaloussein Rezai	ACECR, Iran
Hassan Tavakoli	Guilan University, Iran
Hamed Vahdat-Nejad	University of Isfahan, Iran
Jafar Mansouri	Ferdowsi University of Mashhad, Iran
Seyed Sahand Mohammadi Ziabari	Malek Ashtar University of Technology, Iran
Sritrusta Sukaridhoto	Politeknik Elektronika Negeri Surabaya, Indonesia
Indra Riyanto	Faculty of Engineering Budi Luhur University, Indonesia
Wisnu Widiarto	Sebelas Maret University, Indonesia

Raveendranathan Kalathil Chellappan	College of Engineering, Thiruvananthapuram, India
Sreedharan Pillai Sreelal	Indian Space Research Organization, India
Dhananjay Kumar	Anna University, India
Seshan Srirangarajan	Indian Institute of Technology Delhi, India
Aditi Sharma	MBM Engineering College Jodhpur, India
Abhishek Das	Tripura Central University, India
Durgesh Mishra	Sri Aurobindo Institute of Technology, India
Ravi Subban	Pondicherry University, Pondicherry, India
Krishna Battula	Jawaharlal Nehru Technological University Kakinada, India
Manish Gupta	Hindustan Institute of Technology and Management, Agra, India
Koushik Majumder	West Bengal University of Technology, India
Rajiv Pandey	Amity University, Lucknow Campus, India
Priya Ranjan	Amity University, India
Arun Sharma	Indira Gandhi Delhi Technical University for Women, India
Poorani Shivkumar	KAHE, India
Urmila Shrawankar	RTM Nagpur University, India
Sanjay Singh	Manipal Institute of Technology, India
Venkateshwara Prasad Tangirala	Godavari Institute of Engineering and Technology, India
Vishnu Srivastava	CSIR-Central Electronics Engineering Research Institute Pilani, India
Sunil Kumar Kopparapu	Tata Consultancy Services, India
Mukesh Taneja	Cisco Systems, India
Vyshnavi Ramesh	Independent Researcher, India
Nisheeth Joshi	Banasthali University, India
Rajiv Singh	Banasthali University, India
Binod Kumar	JSPM's Jayawant Institute of Computer Applications, Pune, India
Annappa B.	National Institute of Technology Karnataka, Surathkal, India
Vivek Sehgal	Jaypee University of Information Technology, India
Ayan Mondal	Indian Institute of Technology Kharagpur, India
Ranjana Rajnish	Amity University, Lucknow, India
Karthik Srinivasan	Philips, India
Naveen Aggarwal	Panjab University, India
Shajith Ali	SSN College of Engineering, Chennai, India
Vivek Singh Bhadouria	National Institute of Technology, Agartala, India
Deepshikha Bhargava	Amity University Rajasthan, India
Dinesh Bhatia	Biomedical Engineering Department, North Eastern Hill University, India

Siddhartha Bhattacharyya	RCC Institute of Information Technology, India
Hitesh Bheda	RK University, India
Monowar Bhuyan	Kaziranga University, India
Mantosh Biswas	National Institute of Technology, Kurukshetra, India
Samit Biswas	Indian Institute of Engineering Science and Technology, Shibpur, India
Rushikesh Borse	University of Pune, India
Swati Chandé	International School of Informatics and Management, India
Vinay Chandna	Jaipur Engineering College and Research Centre, Jaipur, India
Vinod Chandra S.S.	University of Kerala, India
D. Chaturvedi	Dayalbagh Educational Institute, India
Mayank Chaturvedi	Graphic Era University, India
Anirban Chowdhury	MIT Institute of Design, India
Shanmugapriya D.	Avinashilingam Institute for Home Science and Higher Education for Women, India
Ranjan Das	Indian Institute of Technology Ropar, India
Vivek Deshpande	University of Poona, India
Durairaj Devaraj	Kalasalingam University, India
Chitra Dhawale	Amravati University, India
Arvind Dhingra	GNDEC, Ludhiana, India
Mahendra Dixit	SDMCET, India
Subhash Dubey	Govt College of Engineering and Technology Jammu-J&K, India
Nitul Dutta	MEF Group of Institutions, Rajkot, India
Paramartha Dutta	Visva-Bharati University, India
Omid Mahdi Ebadati E.	Hamdard University, India
Ravi G.	Sona College of Technology, India
Niketa Gandhi	University of Mumbai, India
G. Ganesan	Adikavi Nannaya University, India
Rajeev Gupta	Rajasthan Technical University, India
Amudha J.	Amrita Vishwa Vidyapeetham, India
J. Vimala Jayakumar	Alagappa University, India
Satishkumar Joshi	The Maharaja Sayajirao University of Baroda, India
Shriram K. Vasudevan	Amrita University, India
Triveni Keskar	Ramrao Adik Institute of Technology, Navi Mumbai, India
Harish Kumar	King Khalid University, India
Manoj Kumar	Guru Gobind Singh Indraprastha University, New delhi, India
Naresh Kumar	GGSIPIU, India

Rakesh Kumar	National Institute of Technical Teachers Training and Research, India
Vishal Kumar	Bipin Tripathi Kumaon Institute of Technology, Dwarahat, India
Anirban Kundu	Netaji Subhash Engineering College, India
Ashish Mani	Amity University Uttar Pradesh, India
Rajeev Mathur	Geetanjali Instt of Tech Studies, Udaipur, India
Deepak Mishra	IIST, India
Sachin Mishra	Amity School of Engineering and Technology, Amity University, India
Kirit Modi	Ganpat University, India
Rana Mukherji	The ICFAI University Jaipur India, India
Ravibabu Mulaveesala	Indian Institute of Technology Ropar, India
Nimushakavi Murti sarma	JNT University Hyderabad, Hyderabad, India
Madhu Nair	University of Kerala, India
Divya Nalla	Nalla Malla Reddy Engineering College, India
Durgesh Nandan	Jaypee University of Engineering and Technology, India
Asoke Nath	St. Xavier's College, India
Jisha Panackal	Vidya Academy of Science and Technology, India
Shashikant Patil	SVKM's NMIMS Mumbai India, India
Shashikant Patil	SVKM NMIMS Mumbai India, India
Karantharaj Porkumaran	Dr NGP Institute of Technology, India
V. Prem Prakash	Dayalbagh Educational Institute, India
Zaheeruddin	Jamia Millia Islamia, New Delhi, India
Tushar Ratanpara	Dharmsinh Desai University, India
Fathima Rawoof	K S School of Engineering and Management, Bangalore, India
Rahul Roy	Machine Intelligence Unit, India
Sheetal Sahu	Gandhi PR College, BU, Bhopal, India
Ashish Saini	Dayalbagh Educational Institute, India
Padmanabhan Sanjeevikumar	Ohm Technologies, India
Sayantam Sarkar	Vijaya Vittala Institute of Technology, India
Kandasamy Selvaradjou	Pondicherry Engineering College, India
Shivraj Sharma	Rajasthan Technical University, India
Amit Shrivastava	JECRC University, India
Shishir Shukla	Amity University, India
Sarbjeet Singh	Panjab University, Chandigarh, India
Muthukumar Subramanian	Indian Institute of Information Technology, Tamil Nadu, India
G.A. Shanmugha Sundaram	Amrita Vishwa Vidyapeetham University, India
Nagender Suryadevara	Geethanjali College of Engineering and Technology, India

Prasheel Suryawanshi	MIT Academy of Engineering, Alandi (D), Pune, India
Ayush Swarnkar	Rungta College of Engineering and Technology, Bhilai, India
Rajine Swetha	RV College of Engineering, India
Ragunathan Thirumalaisamy	Indian Institute of Design Manufacturing, Kurnool, India
Manish Tiwari	Manipal University, Jaipur, India
Ashish Urkude	University of Petroleum and Energy Studies, India
Seema Verma	University Banasthali Vidyapith, India
Subhasis Bhattacharjee	Synopsys Pvt. Ltd., India
Chandrsekaran Chandramouli	Indian Institute of Technology Bombay, India
Pethuru Raj Chelliah	IBM, India
Abhijeet Khandagale	G.H. Rasoni College of Engineering, India
Anbalagan Thangavel	Robert Bosch Engineering and Business Solution, India
Mohd Asim Aftab	Jamia Millia Islamia, India
Lokesh Garg	Malaviya National Institute of Technology, India
Prakash Kumar	UCE Kota, India
Rajesh Pandey	IIITA, India
Md Sarwar	Jamia Millia Islamia, India
Snigdha Shakya	CEERI, Pilani, India
Deepika Shukla	PDPM IITDM Jabalpur, India
Robert Szabolcsi	Óbuda University, Hungary
Haris Psillakis	National Technical University of Athens, Greece
Dimitrios Stratogiannis	National Technical University of Athens, Greece
Vasileios Baousis	University of Athens, Greece
Tharrenos Bratitsis	University of Western Macedonia, Greece
Konstantinos Giannakis	Ionian University, Greece
Sotiris Karachontzitis	University of Patras, Greece
Stefanos Kollias	NTUA, Greece
Sotiris Kotsiantis	University of Patras, Greece
Ioannis Moscholios	University of Peloponnese, Greece
Abdul-Rahman Ahmed	Kwame Nkrumah University of Science and Technology, Ghana
Otthein Herzog	University of Bremen, Germany
Frank Koussen	RWTH Aachen, Germany
Ilka Miloucheva	Media Applications Research, Germany
Torsten Strasser	University of Tuebingen, Germany
Matthias Vodel	Chemnitz University of Technology, Germany
Munir Georges	Intel, Germany
Sahbi Baccar	CESI Rouen, France
Pascal Lorenz	University of Haute Alsace, France
Farid Naït-Abdesselam	Paris Descartes University, France

Paul Honeine	Université de Rouen, France
Kester Quist-Aphetsi	University of Brest France, France
Mohamed Ba khouya	University of Technology of Belfort-Montbeliard, France
Selma Boumerdassi	Conservatoire National des Arts et Métiers, France
Salah Bourennane	Ecole Centrale Marseille, France
Antoine Doucet	University of La Rochelle, France
Said Hoceini	UPEC, University Paris-Est Creteil Val de Marne, France
Francine Krief	University of Bordeaux, France
Alain Lambert	University of Paris Sud, France
Abdallah Makhoul	University of Franche-Comté, France
Pierre Melchior	Bordeaux-INP/ENSEIRB-MATMECA, France
Thierry Monteil	LAAS-CNRS, University of Toulouse, France
Amir Nakib	University Paris East, France
Gopalasingham Aravinthan	Nokia Bell Labs, France
Md Sahidullah	University of Eastern Finland, Finland
Taneli Riihonen	Aalto University School of Electrical Engineering, Finland
Yar Mughal	University of Tartu, Estonia
Mare Koit	University of Tartu, Estonia
Ehab El-Shazly	Kyushu University & EJUST, Egypt
Lamiaa Elrefaei	Benha University, Egypt
Maki Habib	The American University in Cairo, Egypt
Sherief Hashima	Engineering dept, Nuclear Research Center, EAEA, Cairo, Egypt
Mohamed Moharam	Misr University For Science and Technology, Egypt
Ivo Bukovsky	Czech Technical University in Prague, Czech Republic
Tomas Vogeltanz	Tomas Bata University in Zlin, Czech Republic
Frantisek Zboril	Brno University of Technology, Czech Republic
Kamil Dimililer	Near East University, Cyprus
George Dekoulis	Aerospace Engineering Institute, Cyprus
Josip Music	University of Split, Croatia
Jovana Zoroja	University of Zagreb, Croatia
Ahmed Elmisery	Universidad Técnica Federico Santa María, Chile
Roghoyeh Salmeh	Ph. D., P. Eng. PMP, SM IEEE, FEC FGC (Hon.), Canada
Ljiljana Trajkovic	Simon Fraser University, Canada
Salah Benbrahim	Ecole Polytechnique, Canada
Belloulata Kamel	Université de Sherbrooke, Canada
Adel Sharaf	University of New Brunswick, Canada
Abhijit Sinha	AUG Signals, Canada

Zakia Asad	University of Toronto, Canada
Karaputugala Madushan Thilina	University of Manitoba, Canada
Rossitza Goleva	Technical University of Sofia, Bulgaria
Kiril Alexiev	IICT-Bulgarian Academy of Sciences, Bulgaria
Vania Estrela	Universidade Federal Fluminense, Brazil
Lucio Agostinho	Federal University of Technology-Campus Dois Vizinhos, Brazil
Rodrigo Campos Bortoletto	São Paulo Federal Institute of Education, Science and Technology, Brazil
Wagner Botelho	UFABC, Brazil
Raphael Gomes	Instituto Federal de Goiás-IFG, Brazil
Felipe Henriques	Celso Suckow da Foseca Federal Center of Technological Education-CEFET/RJ, Brazil
Marcel Wagner	University of São Paulo, Brazil
A.F.M. Sajidul Qadir	Samsung R&D Institute-Bangladesh, Bangladesh
Hussain Mohammed Dipu Kabir	The Hong Kong University of Science and Technology, Bangladesh
Ahmed Jameel	Ahlia University, Bahrain
Daniel Watzenig	Graz University of Technology, Austria
Artemios Voyiatzis	SBA Research, Austria
Sasan Adibi	Deakin University, Australia
Waail Al-waely	Griffith University/School of Engineering, Australia
Narottam Das	University of Southern Queensland, Australia
Saeid Nahavandi	Deakin University, Australia
Ligang Zhang	Queensland University of Technology, Australia
Esteban Mocskos	University of Buenos Aires, Argentina
Nour EL Yakine Kouba	University of Sciences and Technology Houari Boumediene, Algeria
Samir Ladaci	National Polytechnic School of Constantine, Algeria
Mohand Lagha	Aeronautical Sciences Laboratory, Algeria
Lahecène Mitiche	University of Djelfa, Algeria
Amad Mourad	University of Bejaia, Algeria
Mohammed Saaidia	University of Souk Ahras, Algeria, Algeria
Amel Serrat	USTO MB, Algeria

Contents

Analysis of Link Prediction in Directed and Weighted Social Network Structure	1
Salam Jayachitra Devi and Buddha Singh	
An Energy-Efficient Fuzzy Based Data Fusion and Tree Based Clustering Algorithm for Wireless Sensor Networks	14
Veeramuthu Venkatesh, Pethuru Raj, and P. Balakrishnan	
Biologically-Inspired Foraging Decision Making in Distributed Cognitive Radio Networks	28
Olukayode A. Oki, Thomas O. Olwal, Pragasen Mudali, and Matthew Adigun	
Efficient Algorithms for Hotspot Problem in Wireless Sensor Networks: Gravitational Search Algorithm	41
Srikanth Jannu, Suresh Dara, Katha Kishor Kumar, and Sabitha Bandari	
Web Service Recommendation Based on Semantic Analysis of Web Service Specification and Enhanced Collaborative Filtering	54
S. Subbulakshmi, K. Ramar, Ameena Shaji, and Parvathy Prakash	
Performance Comparison of Apache Spark and Hadoop Based Large Scale Content Based Recommender System	66
Saravanan S., Karthick K.E., Ashwin Balaji, and Anand Sajith	
Performance Evaluation of AODV Routing Protocol for Free Space Optical Mobile Ad-Hoc Networks	74
Salma Fauzia and Kaleem Fatima	
Combination of Fuzzy Logic Digital Image Watermarking and Advanced Encryption Technique for Security and Authentication of Cheque Image	84
Sudhanshu Suhas Gonge and Ashok Ghatol	

Analysis of AES-GCM Cipher Suites in TLS	102
B. Arunkumar and G. Kousalya	
Data Classification Using Machine Learning Approach	112
Shekhar Pandey, Supriya M, and Abhilash Shrivastava	
Topic Modeling for Unsupervised Concept Extraction and Document Ranking	123
V.S. Anoop, S. Asharaf, and P. Deepak	
Markov Chain Monte Carlo Methods and Evolutionary Algorithms for Automatic Feature Selection from Legal Documents	136
S. Pudaruth, K.M.S. Soyjaudah, and R.P. Gunpath	
An Adaptive Soft Set Based Diagnostic Risk Prediction System	149
Terry Jacob Mathew, Elizabeth Sherly, and José Carlos R. Alcantud	
Weighted Bipartite Graph Model for Recommender System Using Entropy Based Similarity Measure	163
Punam Bedi, Anjali Gautam, Saumya Bansal, and Deepika Bhatia	
Automated Quiz Generator	174
Amit Bongir, Vahida Attar, and Ramanand Janardhanan	
Temporal Modelling of Bug Numbers of Open Source Software Applications Using LSTM	189
Jayadeep Pati, Krishnkant Swarnkar, Gourav Dhakad, and K.K. Shukla	
Direct Demodulator for Amplitude Modulated Signals Using Artificial Neural Network	204
Vineetha K.V. and Dhanesh G. Kurup	
Real-Time Detection of Atrial Fibrillation from Short Time Single Lead ECG Traces Using Recurrent Neural Networks	212
V.G. Sujadevi, K.P. Soman, and R. Vinayakumar	
StyloLIT: Stylometry and Location Indicative Terms Based Geographic Location Estimation Using Convolutional Neural Networks	222
K. Surendran, O.P. Harilal, P. Hrudya, and Poornachandaran Prabakaran	
EMG Pattern Classification Using Neural Networks	232
Tanmay Gupta, Jyoti Yadav, Shubham Chaudhary, and Utkarsh Agarwal	
Crime Against Women: A State Level Analysis Using a Hierarchical and K-Means Clustering Techniques	243
Ayushi Dhawan and M.G. Deepika	
Zero Pronouns and Their Resolution in Sanskrit Texts	255
Madhav Gopal and Girish Nath Jha	

Semantic Analysis Using Pairwise Sentence Comparison with Word Embeddings 268
 Vijay Krishna Menon, Sabdhi M., Harikumar K., and Soman K.P.

Illuminant Color Inconsistency as a Powerful Clue for Detecting Digital Image Forgery: A Survey 279
 Divya S. Vidyadharan and Sabu M. Thampi

A Fast, Block Based, Copy-Move Forgery Detection Approach Using Image Gradient and Modified K-Means 298
 V. Hajihashemi and A. Alavi Gharahbagh

OR Operation Based Deterministic Extended Visual Cryptography Using Complementary Cover Images 308
 K. Praveen, G. Indhu, and M. Sethumadhavan

Empirical Comparison of Different Key Frame Extraction Approaches with Differential Evolution Based Algorithms 317
 Kevin Thomas Abraham, Manikandan Ashwin, Darshak Sundar, Tharic Ashoor, and Gurusamy Jeyakumar

Breast Cancer Diagnosis and Prognosis Using Machine Learning Techniques 327
 Sunil Suresh Shastri, Priyanka C. Nair, Deepa Gupta, Ravi C. Nayar, Raghavendra Rao, and Amritanshu Ram

Performance Assessment Framework for Computational Models of Visual Attention 345
 Bharathi Murugaraj and J. Amudha

Structural Matching of Control Points Using V-D-L-A Approach for MLS Based Registration of Brain MRI/CT Images and Image Graph Construction Using Minimum Radial Distance 356
 Hema P. Menon and A.S. Nitheesh

An Interactive and Intelligent Tool for Circuit Component Recognition Through Virtual Reality 370
 Shriram K. Vasudevan, S.N. Abhishek, N.K. Keerthana, Rajan Priyanka, A. Aravinth, and M. Divya

FPGA-Based Heavy-Ion Beam Trajectory Estimation and Control for Superconducting RF Cavity Resonator Applications 380
 B. Christopher, S. Kiruthika, S. Lakshmi, R. Mugunth Krishnan, and G.A. Shanmugha Sundaram

PAM4-Based RADAR Counter-Measures in Hostile Environments 390
 S. Srivatsa and G.A. Shanmugha Sundaram

Multi-criteria Decision Making on Lattice Ordered Multisets 401
 V.S. Anusuya Ilamathi and J. Vimala

Author Index 417

Analysis of Link Prediction in Directed and Weighted Social Network Structure

Salam Jayachitra Devi^(✉) and Buddha Singh

School of Computer and Systems Sciences, Jawaharlal Nehru University,
New Delhi 11067, India

jayachitra.salam@gmail.com, b.singh.jnu@gmail.com

Abstract. The main aim of this paper is to develop algorithms for link prediction based on directed and weighted social network structure. In this paper, the four algorithms such as Modified Common Neighbor (MCN), Modified Jaccard's Coefficient (MJC), Modified Adamic Adar (MAA) and Modified Preferential Attachment (MPA) has been proposed which is suitable for directed and weighted networks. In our proposed algorithms, the degree of nodes and weightage of each link has been considered. The weightage of each link is assigned using random function. The Modified Common Neighbor (MCN), Modified Jaccard's Coefficient (MJC), and Modified Adamic Adar (MAA) algorithms are based on an existing Common Neighbor algorithm. The Modified Preferential Attachment (MPA) algorithm depends on the degree of the nodes. The comparative analysis of our proposed algorithms and existing algorithms is performed based on area under receiver operating characteristic values (AUC values), considering different observed links. According to the experimental analysis, it may be concluded that our proposed algorithms provide better performances in comparison to the existing algorithms. Modified Common Neighbor and Modified Adamic Adar results in highest AUC value when twitter dataset and amazon dataset is considered. The proposed algorithms will be applicable in different directed and weighted social network structure for prediction of links between the users.

Keywords: Social network · Link prediction · Directed and weighted network · Network structure · AUC

1 Introduction

Social Network provides a platform for representatives of sociology, biology, economics, communications science, human geography, etc. [1, 2]. Social network consists of several individuals and the interaction is carried out among the individuals. The interaction among them is denoted by links between the nodes [3, 4]. Prediction of new links among the individuals is an emerging problem in social network analysis. Social network structure plays a different role in data transmission, in political campaigns, in spreading of disease and in many other fields. Therefore, the study of network structure and their properties have become the most emerging topics in many branches of science. Some of the tasks that can be applied to network data are like classification of individuals, prediction of links between those individuals, etc. [5].

Social networks are dynamic in nature as it results in frequent addition and removal of individuals and links among the individuals in the network [6–9]. Link prediction is very useful in practical application, namely in biological network as it can save experimental time and cost [10]. Prediction of link is mainly related to web mining and graph theory [11, 12]. Link prediction is the most important research field in this present situation [13]. Link prediction exploits the information of the network in order to predict the relationships that are highly probable to be formed in the future. It predicts the link between the individuals according to the topological information of the network instead of referring to the attributes of the individuals. The main reason is that the actual connection of the network is based on the topological information. Therefore, it concludes that both attributes of individuals and the observed links can be used mainly for the link prediction in the future [4, 6, 9, 14, 15]. Some of the existing methods focus on the number of connections with the other individuals. Higher the number of connections of individuals with different people, the higher the likelihood to set up link in future. The real world networks are dynamic in nature, hence link prediction is considered to be one of the most challenging tasks. In typical social network structure, namely Co-authorship network and friendship network is a weighted network structure. In weighted network structure links are classified as strong links and weak links [16–18]. For link prediction, it is natural to consider the weights of the links. The weights represent the type of interaction between the individuals, whether it is a weak connection or strong connection. Likewise, in directed network the direction of edges represent the flow of information from one individual to another. So this motivates us to propose link prediction methods based on weighted and directed network structure and investigate the performance with the existing methods.

From the literature survey, we come to know that several researchers have developed various algorithms of link prediction. Some of the existing algorithms are classified according to the network structure. Link predictions based on the directed network have been proposed by various researchers. This method is mainly used in directed network to identify the missing and spurious interactions [1]. We have also come across link prediction methods based on clustering information from the literature studies [4]. Link prediction method based on hypergraph have been proposed, in which social network have been model as hypergraph and prediction of higher order link without any loss of information is carried out. But such methods are applicable only with hypergraph model network [5]. According to Liben-Nowell and Kleinberg, analysis of link prediction based on the proximity of nodes was developed. Large co-authorship datasets are used to carry out the experiments. Topological information of the network is used for future interaction. Still improvement in the efficiency on a large network based on proximity of node methods is needed [6].

Supervised learning methods for prediction of link in terms of various performance measures like accuracy, F-values, precision-recall etc. have been proposed using different classification algorithms on network datasets to predict the performances, but they consider only co-authorship network [7]. Researchers also works on directed network for link prediction and the uses of link prediction in microblog. They proposed effective and efficient methods for prediction of link consisting of three steps. First, for a target node similar nodes are identified. Then identify the candidates, which are linked by the similar nodes. Finally, candidates are rank according to the weight

schemes. The experiments to verify the accuracy of the proposed methods is carried out using real data of microblog [10]. Methods based on time series using node similarity measures have been proposed. In these methods, they compute different scores of node similarities and further extend to weighted version using time series. ARIMA, time series model is used to predict the future link [15]. Link prediction methods based on the centrality of common neighbor nodes have also been proposed. Three types of centrality measures are used: degree centrality, betweenness centrality and closeness centrality. In spite of these, theory of weak ties is considered in order to improve the accuracy of link prediction [16].

Researchers also developed algorithms that influence the weight of the edges. It has been observed that the results of different weighting method vary according to different datasets and different methods of link prediction [18]. A recommendation algorithm for link prediction have been proposed. The author used bipartite network and in recommendation process domain knowledge is incorporated along with topological property. But these algorithms are design on customer-product bipartite network [19]. Researchers also works on link prediction for the Bipartite Social Network, which include users of different roles in the implementation of structural holes. These structural holes are useful in increasing accuracy of link prediction [20].

From the survey of an existing algorithm, we have come to know that algorithm exists for different network structure, but still we have not come across algorithms related to weighted and directed network model. It will be quite difficult to extend the method of undirected network structure to the method of directed and weighted network structure. We can consider this task to be a challenging task for the researcher as it will need a more knowledge of understanding the direction of the flow of information and weightage of links, in order to implement the new algorithm. So, this motivates us to proposed methods for directed and weighted network structure.

The main objective of this paper is, we consider the direction of each link such as indegree and outdegree and weight of each outgoing and incoming link. In order to verify the performance of our proposed methods, we consider some of the undirected and unweighted network structure. Then, we extended the existing methods of undirected and unweighted network structure to the methods of directed and weighted network structure. Finally, we determine the AUC of our proposed methods for predicting future links.

The paper consists of different sections and it is organized as follows: Sect. 2 describes about the problem description and evaluation metric. Sections 3 and 4 present our proposed methods and the evaluation of our proposed methods in directed and weighted network structure. Sections 5 and 6 gives the simulation parameter and the performance analysis. Finally, followed by conclusions and future work.

2 Network Modeling Prediction

Consider a network model in which links are assigned with weights where it represents the strength of the link and the direction of flow of the information. These networks can generally be represented as a graph which include nodes, directed edges and weights of the edges. Given $G(V, E, W)$ where V gives the set of nodes, E represents the set of

edges and W denotes the strength of all the links in the network model. In this network model connection to itself and multiple linking is not acceptable. The main aim of link prediction is to determine the rank of all the non-existing links and the link with a higher rank will be more probable to link in future. As we are considering a directed network model, all possible edge combinations can be predicted by using

$$|V|(|V| - 1) \quad (1)$$

The non-existing edges can be determined by using

$$|V|(|V| - 1) - E \quad (2)$$

But, in an undirected network model the total number of all the possible edge combination [12] is predicted as

$$|V|(|V| - 1)/2 \quad (3)$$

The performance of the algorithm is determined from the area under the receiver operating characteristic curve (AUC). The evaluation of the performance can be done by using the whole list and from the list the highest score value is chosen and the corresponding nodes which results highest score value will establish the new link. So, the AUC can be determined by using the following expression

$$AUC = \frac{n' + 0.5n''}{n'''} \quad (4)$$

Where n' represents missing links which have higher scores as compared to unconnected links and n'' represents how many times they have similar scores. Finally, n''' gives the total number times a pair of links is picked up randomly from a set of missing links and unconnected links. The prediction algorithm works better if it results to higher AUC value. So, higher the AUC value, better the performance of the algorithm.

3 Proposed Method

We first considered some of the existing methods for link prediction and later present the extension that has been made to predict link in directed and weighted network model.

(1) Modified Common Neighbor (MCN)

The main idea used in the common neighbor similarity measures is that in a given undirected network model if any two nodes have a more common friend as compare to others they are likely to form connections in the future. Consider that x and y are the two nodes, $\Gamma(x)$ and $\Gamma(y)$ represents the neighbors of node x and y [20, 21]. The numerical expression is as follows

$$common\ neighbor = |\Gamma(x) \cap \Gamma(y)| \quad (5)$$

Where the generalized representation of determining the neighbor of a node in an undirected network model is $\Gamma(i) = \{k | (i, k) \in E \vee (k, i) \in E\}$. When this method has been extended to directed plus weighted network model, we consider the in degree as well as out degree for every node and also the weights corresponding to those nodes. Generally in degree is represented as $\Gamma_{indegree}(i) = \{j | (j, i) \in E\}$ and outdegree is represented as $\Gamma_{outdegree}(i) = \{j | (i, j) \in E\}$ where i and j represent the node in the network [17]. The weight of a particular link can be determined by using $weight(i, j) = \{w(i, j) | (i, j) \in E \text{ where } (i, j) \neq (j, i)\}$. So, by using this entire idea we can extend the common neighbor similarity measures for directed and weighted network model. The expression of the extended measures is as follows

$$MCN = \sum_{z \in \Gamma_{outdegree}(u) \cap \Gamma_{indegree}(v)} (w(u, z) + w(z, v)) / n \quad (6)$$

Where z is a set of intersection of out-degree of node u and in-degree of node v . Here n represent the highest weights of links. The expression above clearly indicates that the more nodes the set z have the weightage will be more and higher the weightage then higher the probability to connect the two nodes. Considering the real time example, like in Facebook, twitter etc. the more weightage of two individuals have, the higher the probability to have a connection between the two individuals in future.

(2) Modified Adamic Adar (MAA)

This method used the common neighbor method of an individuals and neighbors of those common neighbors. In this method the pair of nodes which have high common neighbor will results to less score and pair of nodes which have a less common neighbor will result to a high score. So, the prediction of link will be done by choosing the pair of nodes which gives less score so that the concept of common neighbor will not be changed [3, 20]. The expression of Adamic Adar is represented as follows

$$adamic\ adar = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log|\Gamma(z)|} \quad (7)$$

The pair of nodes which is needed to check the common neighbor, need not be connected. If the pair of nodes are already connected from before then, the linking of those nodes does not bring must benefit.

This method can also be extended for directed plus weighted network model. When we consider the directed plus directed network model first thing we should keep in mind is about the incoming edges and outgoing edges. Like the way that has been mentioned above, we should calculate the in-degree and out-degree of each node in the network. The mathematical expression of this proposed method is as follows

$$MAA = \sum_{z \in \Gamma_{outdegree}(u) \cap \Gamma_{indegree}(v)} ((w(u, z) + w(z, v)) / n) \times \frac{1}{\log\left(\sum_{z \in \Gamma_{outdeg}(u)} w(u, z) \times \sum_{z \in \Gamma_{indeg}(v)} w(z, v)\right)} \quad (8)$$

The idea behind this extended method will be similar to the Adamic Adar method for undirected network model. Higher the number of in-degree and out-degree of a pair of nodes the score will be less. That means, if a pair of nodes has less connection, then they are likely to connect in the future.

(3) Modified Jaccard's Coefficient (MJC)

Jaccard's coefficient is most widely used similarity measure for information retrieval. In Jaccard's coefficient if the number of total connection between any pair of nodes is large then, it results to less score. The pair of nodes which have less number of connection will have a high probability to connect in the future [12, 20]. This method is presented mathematically as follows

$$jaccards\ coefficient = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \quad (9)$$

This undirected Jaccard's coefficient similarity measure can be further extended to directed and weighted network method by considering the incoming edges and outgoing edges separately. The expression of proposed method can be expressed as follows

$$MJC = \frac{\sum_{z \in \Gamma_{outdegree}(u) \cap \Gamma_{indegree}(v)} (w(u, z) + w(z, v)) / n}{\sum_{z \in \Gamma_{outdegree}(u)} w(u, z) + \sum_{z \in \Gamma_{indegree}(v)} w(v, z)} \quad (10)$$

This method is also based on common neighbor method. It is applicable in different directed and weighted network. The AUC calculation of this method will be shown in the next section. According to the AUC calculation, we can predict which method gives the most accurate result.

(4) Modified Preferential Attachment (MPA)

Considering many real world networks, this method will be the most basic method that can be used for the prediction of links in a social network. The main idea behind this method is that if a node is more connected, then, it is likely to receive new links in the future. The nodes which have higher degree are much stronger to grab the links that have been newly added to the network. Finally, Preferential Attachment similarity measure concludes that, the probability of connection of nodes is proportional to the number of neighbors [1, 12]. The method can be expressed as follows

$$preferential\ attachment = |\Gamma(x)| \times |\Gamma(y)| \quad (11)$$

Where $\Gamma(i)$ represents the neighbor of node i . Using feedback structure in directed and weighted network model, this preferential attachment model can be expressed as follows

$$MPA = \left[\sum_{z \in \Gamma_{outdegree}(u)} w(u, z) \right] \times \left[\sum_{z' \in \Gamma_{indegree}(v)} w(z', v) \right] / n \quad (12)$$

The proposed method is based on the feedback structure which consider out degree and in degree. Here in preferential attachment method for two nodes, let's consider as u and v , the sum of weight of the out-degree edge of node u is multiplied by the sum of weight of the in-degree edge of node v . In this method, the prediction of link will be based on the highest score value. Higher the score value, the probability of connection will be high.

The evaluation of all these proposed methods will be discussed in the following section and comparison of all the methods and their accuracy value will be evaluated.

4 Evaluation of our Proposed Methods in Directed and Weighted Network Structure

This section will give the detailed evaluation of our proposed link prediction methods. Let us consider Fig. 1, which is a directed and weighted network structure. The most basic approach of prediction of link is to determine the score of all pairs of node from the given network structure. After the calculation of the score for each pair of nodes, it is inserted into a list and finally the list is arranged in descending order so that the highest score will be available on top of the list. Out of all the score, all the highest value is reinserted in the new list and out of all those value any corresponding pairs of nodes can predict the new connection.

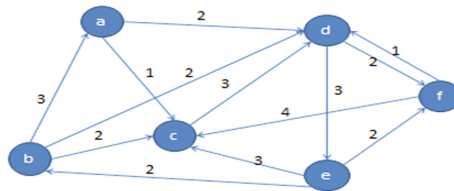


Fig. 1. Weighted and directed graph

First, check the pair of node which is unconnected. If they are connected from before then, it doesn't have much benefit to determine the score. So, we focused on the unconnected pair of node. In Figure node a and f are not connected so we can determine the score by using the four similarity measures. Likewise, we can determine for all the remaining node pairs. The score of Modified Common Neighbor between node a and f is 1. But, when we consider the graph as undirected and unweighted graph the score of node pair a and f is 2 as it calculate the total number of common neighbor between the pair of node. Similarly, for Modified Adamic Adar, Modified Preferential Attachment and Modified Jaccard's Coefficient are introduced in a similar manner.

5 Simulation Parameter

For the evaluation of the proposed algorithms, several simulation parameters are used. First of all, we would like to consider datasets of twitter and amazon which consist of several number of individual nodes in the network required for the evaluative study. The datasets are extracted using NodeXL. This is followed by the number of incoming and outgoing links of each individual node. Probabilistically, weightage for each link is assigned with the help of random function. Table 1 shows the number of nodes and edges are available in the datasets.

Table 1. Datasets

Datasets	No. of nodes	No. of edges
Twitter	55	233
Amazon	291	1450

6 Performance Analysis

This section will present the experimental result that has been evaluated by using Python interpreter 2.7.11. The performance of our proposed algorithms is carried out in terms of AUC of all the algorithms using python program. Observed links are represented as l . In the performance analysis process, we considered a different number of observed links i.e $l = 5$ and $l = 10$. Finally, we have found that our proposed methods are more efficient than existing methods. The following Figure will show the analysis of results obtained.

Table 2. AUC values for proposed algorithms for Twitter dataset

Algorithms	AUC value ($l = 5$)	AUC value ($l = 10$)
MCN	0.8721	0.8538
MJC	0.7234	0.7914
MAA	0.8452	0.7996
MPA	0.8259	0.8094

Table 3. AUC values for proposed algorithms for Amazon datasets

Algorithms	AUC value ($l = 5$)	AUC value ($l = 10$)
MCN	0.8245	0.7857
MJC	0.7849	0.7741
MAA	0.8549	0.8245
MPA	0.8154	0.8049

Table 4. AUC values for existing algorithms for Twitter datasets

Algorithms	AUC value (l = 5)	AUC value (l = 10)
CN	0.7477	0.7431
JC	0.7212	0.7013
AA	0.7216	0.6120
PA	0.8050	0.8056

Table 5. AUC values for existing algorithms for Amazon datasets

Algorithms	AUC value (l = 5)	AUC value (l = 10)
CN	0.7519	0.7372
JC	0.6947	0.7441
AA	0.8824	0.7940
PA	0.8109	0.7856

Table 2 shows the AUC value based of different observed links for all the proposed methods for twitter datasets. MCN gives the highest AUC value for both conditions. Table 3 shows the AUC value based of different observed links for all the proposed methods for Amazon datasets. MAA gives the highest AUC value for both conditions. Table 4 shows the AUC value based of different observed links for all the existing methods for twitter datasets. PA gives the highest AUC value for both conditions. Table 5 shows the AUC value based of different observed links for all the existing methods for Amazon datasets. AA gives the highest AUC value for both conditions.

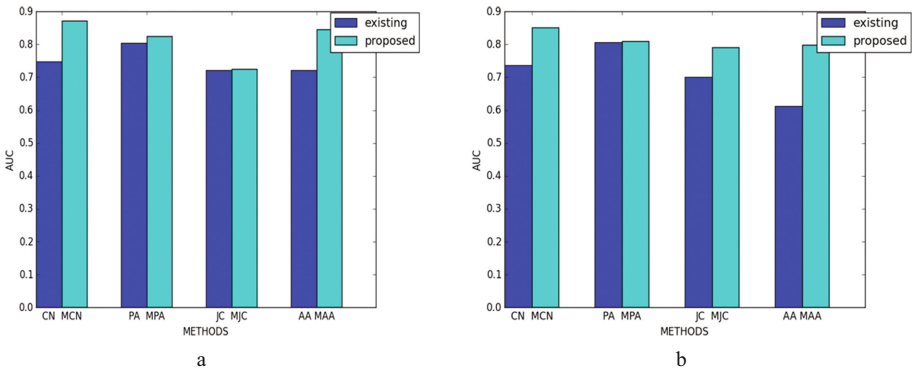


Fig. 2. (a) Comparison of existing algorithms and proposed algorithms with $l = 5$ for Twitter network (b) Comparison of existing algorithms and proposed algorithms with $l = 10$ for Twitter network

In the Fig. 2(a) and (b) show, Common Neighbor (CN), Jaccard's Coefficient (JC), Adamic-Adar (AA), Preferential Attachment (PA) represents the existing algorithm for

undirected and un-weighted network. Modified Common Neighbor (MCN), Modified Jaccard’s Coefficient (MJC), Modified Adamic-Adar (MAA), Modified Preferential Attachment (MPA) represent our proposed algorithms. These modified algorithms are mainly for weighted and directed network. The Figure represents the AUC of our proposed algorithms and existing algorithms for twitter network with an observed links of $l = 5$ and $l = 10$. It is clearly visible that the algorithms for weighted directed network are more accurate and more efficient than the algorithms for un-weighted undirected network. In Fig. 2(a) out of all the proposed algorithms, Modified Common Neighbor gives the highest AUC value and Modified Jaccards Coefficient gives the lowest AUC value. Likewise, in Fig. 2(b) proposed algorithms gives better AUC as compared to existing algorithms.

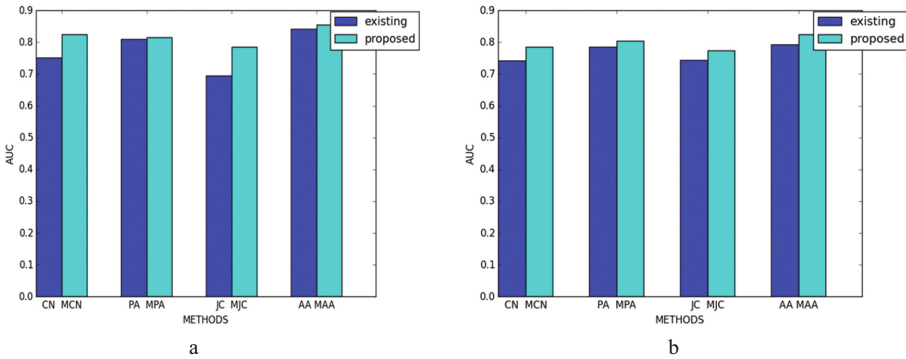


Fig. 3. (a) Comparison of existing algorithms and proposed algorithms with $l = 5$ for amazon network (b) Comparison of existing algorithms and proposed algorithms with $l = 10$ for amazon network

In Fig. 3(a) and (b), the comparison of existing and proposed algorithm regarding amazon network is considered with different observed links of $l = 5$ and $l = 10$. Modified Adamic Adar gives the highest AUC value in both the Figures. And Modified Jaccards Coefficient gives the lowest AUC value.

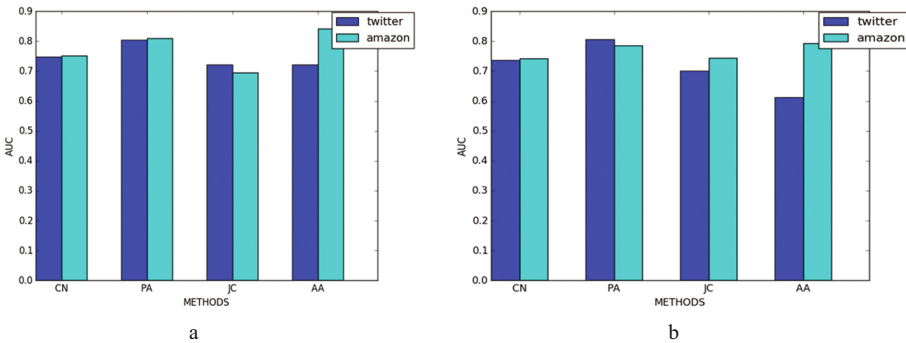


Fig. 4. (a) Comparison of existing algorithms for twitter and amazon network with $l = 5$ (b) Comparison of existing algorithms for twitter and amazon network with $l = 10$

Figure 4(a) and (b) represents the comparison of existing algorithms for Twitter and Amazon network. These Figures will help us to understand the existing algorithms which give the highest AUC and lowest AUC with different observed length of $l = 5$ and $l = 10$. While considering $l = 5$ Adamic Adar of Amazon network gives the highest AUC. And in case of $l = 10$ Preferential Attachment of the Twitter network gives the highest AUC value.

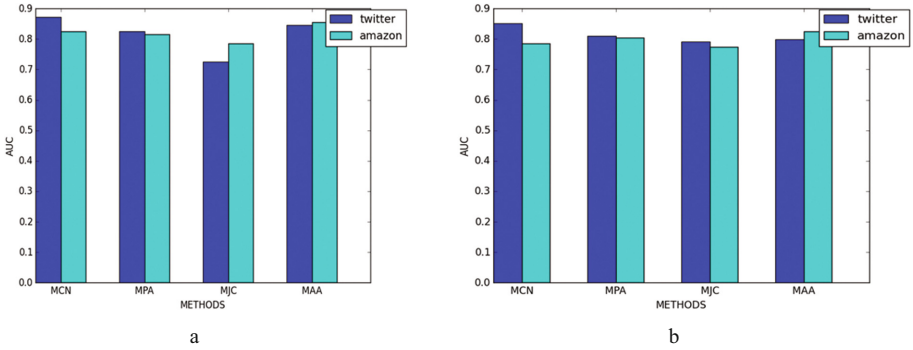


Fig. 5. (a) Comparison of proposed algorithms for twitter and amazon network with $l = 5$ **(b)** Comparison of proposed algorithms for twitter and amazon network with $l = 10$

Figure 5(a) and (b) represents the comparison of proposed algorithms for Twitter and Amazon network. These Figures will help us to understand the proposed algorithm which give the highest AUC and lowest AUC with different observed length of $l = 5$ and $l = 10$. While considering $l = 5$ and $l = 10$ Modified Common Neighbor of Twitter network gives the highest AUC. And Modified Jaccard's Coefficient of both Twitter and Amazon network gives the lowest AUC.

7 Conclusion and Future Work

In social network, link prediction play vital role in understanding the dynamic behavior of various types of complex structure. The existing methods are not applicable in the directed and weighted network. Our proposed schemes are applicable for directed and weighted network. The weightage of the link defines the frequent interactions of the nodes in the network. The weightage of the link increase with the increase in interaction between nodes and decrease if interaction doesn't occur among the nodes. But, our proposed methods initially assigned the weightage of link in the network using random function. Further, the weightage of the links depends on the number of occurrence or non-occurrence of interaction among the nodes. The proposed methods of link prediction are analyzed using Twitter and Amazon datasets by considering the different observed link of $l = 5$ and $l = 10$. We have developed four algorithms Modified Common Neighbor(MCN), Modified Jaccard's Coefficient(MJC), Modified Adamic Adar(MAA) and Modified Preferential Attachment(MPA) respectively.

All these algorithms are applicable in the directed and weighted network. The proposed MCN algorithm computes the weights of links for each pair of unconnected nodes and also compute common neighbor of nodes in the network. Another proposed MJC algorithm is determined as directed and weighted common neighbor divided by the total sum of weight of each outgoing link and incoming link. The modified MAA algorithm used the concept of MCN and it is further extended by multiplying with inverse logarithmic of the sum of each out-degree and in-degree nodes in the network. The MPA algorithm is developed and this is based on the multiplication of the sum of weights corresponding to out degree and sum of weights corresponding to in degree divided by the highest weights of the links.

The proposed methods are applicable in different types of network structure and the performance of our methods improved if we increase the total number of nodes in the network. The performance of proposed algorithms MCN, MJC, MAA and MPA are evaluated with respect to existing algorithms. The results obtained in simulation shows our proposed methods outperform in terms of efficiency and scalability.

As direction for future study, we would like to develop link prediction methods, which will be applicable in different real time social network and will also be able to predict link in the hybrid complex network. In future, our proposed work can be applied to a larger experimental dataset. Also, methods can be extended by considering time domain. More attributes can be considered that depend on time and those attribute values can also be further evaluated using weights. The structural attributes and descriptive attributes of nodes can be combined and this combination of attributes will be applicable in directed and weighted network for link prediction. We can extend our propose method for community detection where the nodes with high AUC value can be grouped under same community. Our work can be explored further by using the more implicit information of the users as, it may be able to give better performance in future. The proposed methods make suitable for bipartite network structure. The methods will consider the semantic information of the nodes in a social network for analyzing the behavior of complex social network.

Acknowledgment. This work is supported by the Adhoc and Wireless Sensor Lab under School of Computer & Systems Sciences and DST purse of Jawaharlal Nehru University, India.

References

1. Zhang, X., Zhao, C., Wang, X., Yi, D.: Identifying missing and spurious interactions in directed networks. *Int. J. Distrib. Sens. Netw.*, 470–481 (2014)
2. Furht, B.: *Handbook of Social Network Technologies and Applications*. Springer Science & Business Media, New York (2010)
3. Bliss, C.A., Frank, M.R., Danforth, C.M., Dodds, P.S.: An evolutionary algorithm approach to link prediction in dynamic social networks. *J. Comput. Sci.* **5**(5), 750–764 (2014)
4. Li, F., He, J., Huang, G., Zhang, Y., Shi, Y.: Retracted: A clustering-based link prediction method in social networks. *Procedia Comput. Sci.* 432–442 (2014)
5. Li, D., Xu, Z., Li, S., Sun, X.: Link prediction in social networks based on hypergraph. In: 22nd International Conference on World Wide Web, 13 May 2013, pp. 41–42. ACM (2013)

6. Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. *J. Assoc. Inf. Sci. Technol.* **58**(7), 1019–1031 (2007)
7. Al Hasan, M., Chaoji, V., Salem, S., Zaki, M.: Link prediction using supervised learning. In: *SDM06: Workshop on Link Analysis, Counter-Terrorism and Security*, 20 April 2006
8. Javari, A., Jalili, M.: Cluster-based collaborative filtering for sign prediction in social networks with positive and negative links. *ACM Trans. Intell. Syst. Technol.* **5**(2), 24 (2014)
9. Gupta, N., Singh, A.: A novel strategy for link prediction in social networks. *CoNEXT on Student Workshop*, 2 December 2014, pp. 12–14. ACM (2014)
10. Yu, Y., Wang, X.: Link prediction in directed network and its application in microblog. *Math. Prob. Eng.* (2014)
11. Fire, M., Tenenboim-Chekina, L., Puzis, R., Lesser, O., Rokach, L., Elovici, Y.: Computationally efficient link prediction in a variety of social networks. *ACM Trans. Intell. Syst. Technol.* **5**(1), 10 (2013)
12. Lü, L., Zhou, T.: Link prediction in complex networks: a survey. *Physica A* **390**(6), 1150–1170 (2011)
13. Wang, T., Liao, G.: A review of link prediction in social networks. In: *2014 International Conference on Management of e-Commerce and e-Government (ICMeCG)*, 31 Oct 2014, pp. 147–150. IEEE (2014)
14. Murata, T., Moriyasu, S.: Link prediction of social networks based on weighted proximity measures. In: *IEEE/WIC/ACM International Conference on Web Intelligence*, 2 Nov 2007, pp. 85–88. IEEE Computer Society (2007)
15. Güneş, İ., Gündüz-Öğüdücü, Ş., Çataltepe, Z.: Link prediction using time series of neighborhood-based node similarity scores. *Data Min. Knowl. Discov.* **30**(1), 147–180 (2016)
16. Liu, H., Hu, Z., Haddadi, H., Tian, H.: Hidden link prediction based on node centrality and weak ties. *EPL Europhys. Lett.* **101**(1), 18004 (2013)
17. Mengshoel, O.J., Desai, R., Chen, A., Tran, B.: Will we connect again? Machine learning for link prediction in mobile social networks. In: *Eleventh Workshop on Mining and Learning with Graphs* (2013)
18. Sett, N., Singh, S.R., Nandi, S.: Influence of edge weight on node proximity based link prediction methods: an empirical analysis. *Neurocomputing* **172**, 71–83 (2016)
19. Li, J., Zhang, L., Meng, F., Li, F.: Recommendation algorithm based on link prediction and domain knowledge in retail transactions. *Procedia Comput. Sci.* **31**, 875–881 (2014)
20. Xia, S., Dai, B., Lim, E.P., Zhang, Y., Xing, C.: Link prediction for bipartite social networks: the role of structural holes. In: *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 26 Aug 2012, pp. 153–157. IEEE (2012)
21. Gao, F., Musial, K., Cooper, C., Tsoka, S.: Link prediction methods and their accuracy for different social networks and network metrics. *Sci. Program.* (2015)

An Energy-Efficient Fuzzy Based Data Fusion and Tree Based Clustering Algorithm for Wireless Sensor Networks

Veeramuthu Venkatesh¹(✉), Pethuru Raj², and P. Balakrishnan³

¹ School of Computing, SASTRA University, Thanjavur 613401, India
veeramuthuvenkatesh@cse.sastra.edu

² Reliance Jio Cloud Services (JCS), Bangalore 560025, India

³ SCOPE, Department of Analytics, VIT University Vellore Campus,
Vellore, India

Abstract. The realization of any wireless sensor network (WSNs) clearly determined on how the key quality of service (QoS) attributes/non-functional requirements (NFRs) gets accomplished. The well-known issues to be taken into account while designing WSNs for setting up smarter environments are information precision (timeliness and accuracy), data accretion, network latency, and energy efficiency. A wisely employed clustering algorithm for interconnecting sensor nodes can significantly enhance the energy efficiency of WSNs. However, the aspect of clustering involves additional overheads due to cluster head selection and cluster construction. This research work proposes a workaround that utilizes Type-II fuzzy for fusing the data and tree driven clustering algorithm that employs Type-2 fuzzy logic to improve the QoS parameters as well as preserving the power/energy of sensor networks. The primary objectives of the proposed cluster algorithm are two folded. Firstly, it constructs the clusters and chooses the cluster head (CH) by considering the remaining energy in the nodes and its distance from the base station (BS). Secondly, it performs the data fusion which contains meaningful information that has been sensed and captured. An extensive experimental analysis has been done on the proposed FBDF-TBC method by comparing it against its counterparts. The simulation results conclude that the proposed fuzzy-based technique for data fusion and tree-based clustering routing algorithm (FBDF-TBC) outperforms other clustering algorithms and improves the overall network lifetime of WSN from a minimum of 16% to maximum of 76%.

Keywords: Mobile sensor networks · Routing protocol · Network lifetime · Fuzzy based clustering · Delay · Data fusion

1 Introduction

The wireless sensor networks (WSNs) are gaining unprecedented popularity due to the distinct advantages that they bring to the table. That is, sensors emerge as a low-cost alternative for a wider variety of application areas across multiple industry verticals [1]. The application domains of WSNs are growing consistently with researchers unearthing

novel use cases. The well-known application areas are pattern/activity/gesture recognition, environmental monitoring, event-driven applications, self, surroundings and situation awareness, safety, surveillance and security applications, the formation of smart environments such as smarter homes, hotels, and hospitals, and so on. Primarily there are two types of WSNs: homogeneous and heterogeneous. An apostolic structure of several sensors are contemplated in these two varieties of WSNs. The lifetime and the dependability of any sensor networks can be enhanced by the heterogeneity aspect. Heterogeneous sensor networks (HSNs) are highly beneficial because they are very right and relevant to many real-world and life scenarios [2]. However, the task of routing has been a challenging concern in the design of WSNs. However, the earlier research works have come out with a few pioneering routing protocols to reduce the energy consumption by nodes in order to enhance WSN routing [3–6]. The concept of clustering is pronounced as an important factor to substantially increase the lifetime of WSNs as clusters are typically able to significantly decrease energy consumption [7]. Clusters come handy in setting and sustaining scalable sensor networks in order to tackle more data loads. The availability of sensor networks is guaranteed through clustering. Thus the concept of clustering of various participating and contributing nodes is being termed as the most important domain for intense study and research. The proven master and slave concept is doing well in forming sensor networks. There have to be one or more master nodes in order to keep the slave nodes well. If there is any fallout or problem with worker nodes, the master node has to do the necessary corrective actions immediately in order to finish the work started. The master node is typically touted as the cluster head (CH). The cluster head is nominated and designated by all the sensors in the cluster. That can be also decided by network designer. The traditional clustering protocols in WSNs assume that every nodes in the sensor networks remain stuffed with the same amount of energy. This assumption comes in the way of leveraging the extreme benefits of heterogeneous nodes and networks. In order to use the heterogeneity, clustering procedures are mainly classified based on two main benchmarks according to its meta-stability and energy-efficiency factors. The selection of cluster head for deriving energy-efficient networks generally depends on the early energy, residual energy and intermediate energy of the network and the energy depletion rate or the mixture of these parameters. The chosen protocols for clustered HSNs prolong the time intermission before the death of first node. This is called the meta-stability period.

1.1 The Key Contributions of the Proposed Framework Are Foregrounded as Follows

Tree-based clustering algorithm (TBC) for effective cluster formation. The limited battery power of each node is a major factor capable of adversely impacting the lifetime of the entire network. Tree-based [3–7] thereby it is a good ploy for extending the network lifetime. While performing tree-based clustering, there is a need to smartly construct clusters to decrease the communication distance among the sensor nodes.

Type-2 Fuzzy Logic Based Sensor Data Fusion (FBDF) for the Removal of Redundant Information Thereby Improving the Energy. Sensor Data Fusion Removes Any Incorrect and Duplicated Values from Sensors to Increase the Intended QoS. There Are Many Fusion Techniques [8, 9] Available to Perform Data Fusion but Type-2 Fuzzy Logic Provides a Greater Accuracy and Reduces Energy Utilization in Sensing the Environment

Distributed Source Coding for compression. In addition, the compression of data before transmitting greatly reduces the energy consumption by decreasing the number of bits to be transferred. In the case of sensor nodes, the common compression technique such as Huffman is not suitable as it requires an enormous amount of memory and robust processing element capability. Distributed Source Coding (DSC) [10] method by Slepian-Wolf theorem provides a precise restoration of data for twofold associated sources using side data sources. Therefore DSC is decided as the most appropriate compression technique for WSN so as to save energy.

The following part of paper is described as: In Sect. 2, a study on existing methods are described. The framework and the problem summary are briefed in Sect. 3 followed by Sect. 4 that elucidates the proposed framework of fuzzy based data-fusion technique. Section 5 elucidates the brief analysis of the performance of the framework. At last, the conclusion and future enhancements is summarized in Sect. 6.

1.2 Related Work

Firstly, this research work supplies the detailed literature survey on any energy aware in wireless sensor networks and groups them based on their objectives (Cluster formation, Cluster head selection, fault-tolerance, packet delay, Sub-clustering, and compression). Secondly, the literature surveyed clearly identifies the gaps, articulates the objectives of the proposed work and carefully formulate the solution methodologies for these objectives. In [11] proposed a fuzzy-based unequal clustering procedure in WSNs to generate clusters with different sizes and this arrangement addresses the persistent hotspot problem. As a result, the paper claims that their method decreases the intra-cluster functions of the cluster-heads which are close to the base station or have low residual energy. In [12] proposed a new methodology to construct a data gathering with energy efficiency as main motto in wireless mobile sensor networks. Subsequently, the paper concludes that the suggested approach minimizes the delay per round and guarantees an improved throughput, and eliminates minimum coverage cost of any underlying network. In [13] have come out with an architecture for any wireless micro sensor networks which are application-specific protocol that includes low-energy adaptive clustering hierarchy (LEACH) to combine the benefits of energy efficiency and lifetime and media access to attain better network lifetime, response time, and application- comprehended value. In [14] proposed a three fuzzy descriptors method based cluster-head election for wireless sensor networks which is more suitable for medium sized clusters. However, the authors highlight that the articulated model introduces a substantial increase in the network lifetime. In [15] have introduced a deterministic clustering protocol for energy saving which, as per the claim of the paper, reduces processing element overhead cost to automate the sensor network, which are

getting reflected in the system lifetime. Besides, this research work claims that the approach approximates and accentuates an ideal interpretation for sensible energy consumption in an ordered WSNs. In [16] have incorporated predictive cluster head selection using fuzzy-based scheme for WSNs. This research work introduces a parameter called the rate of recurrent communication in addition to the remaining power of any nodes is used to decide the cluster head. The fuzzy logic method evaluates the Cluster Head Selection Probability which is based on the node's previous communication history to decide the Cluster Head. The rate of recurrent communication of sensor node is found to yield better results compared to the earlier works. The network model and the processing Flow of FBDF-TBC routing protocol is described in Figs. 1 and 2 respectively. Further, this section explains energy consumption model, data compression model and the data fusion model of proposed FBDF-TBC. The following assumptions hold good for our proposed architecture.

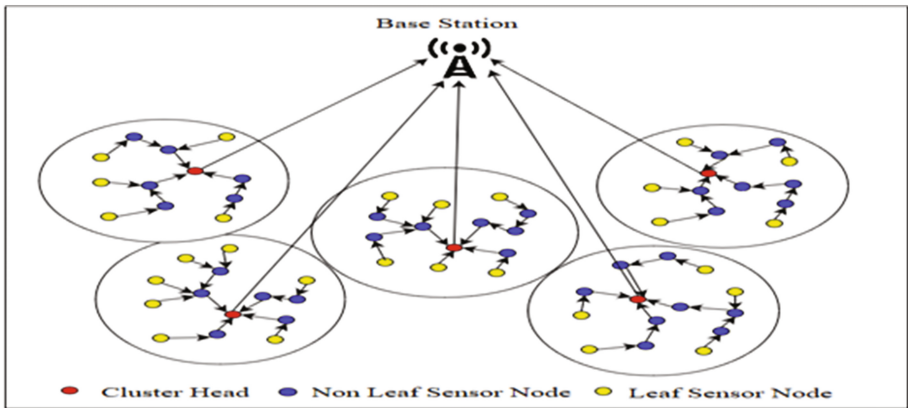


Fig. 1. FBDF-TBC network model

1.3 The Proposed Architectural

The network consists of N nodes equally distributed in a square sensing area and the base station (BS) is far away from the environment that is being sensed.

- BS has unlimited energy resource. The initial battery powers of all the sensor nodes are same initially and the batteries are not rechargeable.
- After the network is deployed, all the sensor nodes and the BS are stationary.
- Each sensor node has same processing and sensing capabilities.

1.4 The Energy Consumption Model

The sensor node majorly comprises of four modules: a power unit; a processing element; a sensing unit; radio frequency transmission unit which consists of an amplifier,

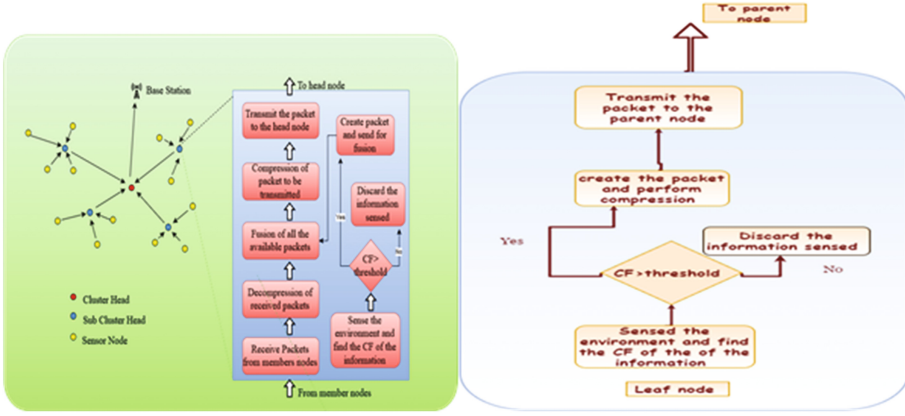


Fig. 2. Process Flow diagram of FBDF-TBC

antenna, and receiver/transmitter circuits. Since the primary objective of this research work is to develop an energy efficient sensor routing protocol that provides accurate sensing information, the energy for transmission and reception is also considered, the energy required to perform data fusion is also taken into consideration. To calculate the transmission energy, the following equations are considered.

$$E_T(q, d) = \begin{cases} q * E_{elec} + q * E_{fs} * d^2, & d < d_{co} \\ q * E_{elec} + q * E_{mp} * d^4, & d > d_{co} \end{cases} \quad (1)$$

Where E_{elec} is the energy consumed by the electrical circuits, q is the size of the packet in bits, d is the space between any twofold nodes, d_{co} is the crossover distance, E_{fs} and E_{mp} are the energies consumed by the amplifiers for distances shorter than d_{co} and distances larger than d_{co} respectively. For receiving a packet of q -bits, the energy consumed is.

$$E_R(q) = q * E_{elec} \quad (2)$$

Hence for a parent node, the energy consumption for a single round

$$E_{parent}(i) = n * E_R(q) + E_F(q) + E_t(q, d(i, j)) + E_S(q) + E_G(q) \quad (3)$$

Where ‘ n ’ represents number of children that have transmitted the packets to the parent node, $E_F(q)$ is the energy for performing data fusion, $E_S(q)$ and $E_G(q)$ are the energies of sensing and generating packets respectively. For the cluster head nodes, the energy consumption is same as $E_{parent}(i)$ except for that $d(i, j)$ is replaced by $d(i, BS)$ where BS is the base station. For the leaf node, the consumption of energy depends on whether the node transmits the data or not. The transmission of the packet is decided by the Type-2 fuzzy logic system. If the sensed data is of greater confidence, then the packet is generated and need to be transmitted otherwise the sensed data is discarded. But for parent nodes, even though the data sensed by it is of lower confidence, it has to

still perform fusion on the packets obtained from its child nodes if any. Thus the energy consumption of a child node is.

$$E_{\text{child}}(\mathbf{j}) = E_t(\mathbf{q}, \mathbf{d}(\mathbf{i}, \mathbf{j})) + E_S(\mathbf{q}) + E_G(\mathbf{q}), \text{ if transmission takes place} \quad (4)$$

$$E_{\text{child}}(\mathbf{j}) = E_S(\mathbf{q}), \text{ if packet is not transmitted} \quad (5)$$

1.4.1 The Compression Model

In WSNs, compression method used to decrease the energy depletion of a node to transfer a packet. By performing compression on the packet to be transmitted, its size is reduced considerably thereby reducing the amount of energy needed to transmit it. However, the most crucial part is to choose an efficient method of compression because the processing capabilities are limited at sensor nodes. This research work opted Distributed Source Coding (DSC) to perform a lossless compression of correlated data values from various sensor nodes.

1.5 The Proposed Routing Methodology

Originally, the inspiration for the development of FBDF-TBC protocol is derived from the extensive analysis of LEACH-C routing algorithm, energy efficient PEGASIS and Type-2 fuzzy logic. The FBDF-TBC routing protocol comprises of the following stages, (a) Clustering and cluster head selection, (b) Cluster Tree formation, (c) Data fusion, (d) Data transmission in the network.

1.5.1 Clustering and Cluster Head Selection

Initially, the configuration (location) and residual energy details are known by BS by transmitting and receiving messages between live nodes and BS. BS separates the entire WSN nodes into five clusters centered on their corresponding proximity. After the cluster creation, a cluster head is selected for each cluster by the BS. Basically, the cluster head alone has the capability to communicate with the BS. Additionally, all the nodes in the cluster transmit to cluster head through a tree based cluster. For example, consider the scenario of 100 live sensor nodes distributed in an area of $100 \times 100 \text{ m}^2$ (refer to Fig. 3). The node at the location (50,175) is the BS and the sensor nodes are distributed into five clusters. The cluster heads are marked with squares and all the other nodes are the non-CH sensor nodes. Further, the member nodes of every cluster are arranged as a minimum spanning tree.

In every iteration, the CH for each cluster is selected by the BS. A sensor node is selected as CH by considering the following two parameters: (a) energy remaining in the sensor node and (b) distance of the node from the base station. The energy E_{average} is used to decide the CH. The node that has residual energy higher E_{average} and paramount cost function is selected as CH for that cluster.

E_{average} Can be calculated as follows:

$$E_{\text{average}} = \frac{\sum_{i=1}^{n\text{Alive}} E_{\text{residual}}(i)}{n\text{Alive}} \quad (6)$$

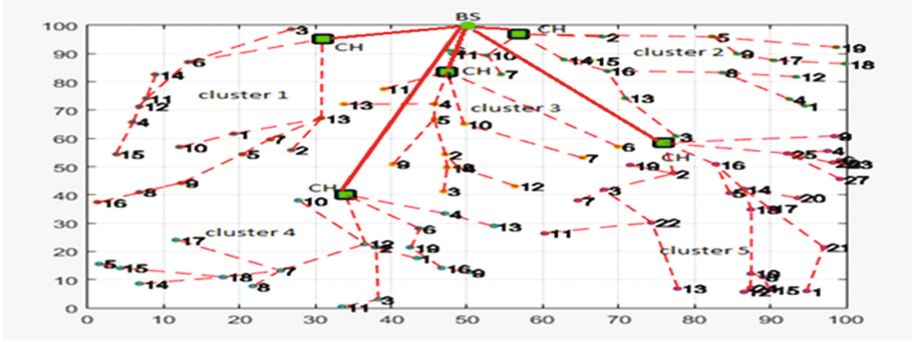


Fig. 3. 100 nodes in a 100 × 100 m² area divided into clusters and a minimum spanning tree is constructed in a round

Where ‘nAlive’ represents the number of live nodes in the cluster and $E_{residual(i)}$ is the remaining energy of the i th node in the cluster. The cost function of a sensor node in a cluster can be calculated as follows:

$$cost(i) = \frac{w_e}{w_d} \times \frac{E_{residual(i)}}{d(i, BS)} \tag{7}$$

Where $d(I, BS)$ is the Euclidean distance between the sensor node and the base station, w_e and w_d are the cost factors of residual energy and distance respectively. The cost factors need to be set appropriately. A Greater w_e value means the available energy of the important node while selecting CH and vice versa.

Algorithm 1. Area Division and Cluster Head Selection

1. INPUT: n-no of cluster required, BS-Base Station node, r- no of round, List of alive nodes
2. OUTPUT: Tree with nodes connected in MST
3. for every node in {Alive Nodes List} doing
 - a. Broadcast NAME packet that contains its ID, residual energy, and Location to BS
4. end for
5. // for each round
6. for x to 1 to r
 - a. Partition the whole network into 'n' clusters with same sub size network
 7. for every cluster(sub-network) in the network do
 - a. Compute the mean energy and Choose the cluster head that the maximum cost function
 8. end for
9. go to Algorithm 2 // Cluster Tree formation phase (MST)//

1.5.2 The Cluster Tree Formation

In this phase, the sensor nodes belonging to respective cluster that are constructed into a minimum spanning tree using Prim's algorithm in such a way that each tree has a minimum sum of weights.

Algorithm 2. Cluster Tree Formation

```

1. weight[CH]=0
2. weight[other nodes]=MAXIMUM
3. {TREE [BS]}=BS
4. while { Alive Nodes List}!= 0} do
    a. Find out a node i in { Alive Nodes List }, where Weight [node i] is minimum
    b. last=node i
       for every node j in { Alive Nodes List } do
           i. if (node j != last) then
               1. if (Weight [node j] > d (node j, last)) then
                   a. weight [node j] = d (node j, last)
                   b. {TREE [node j]}=last
               2. end if
           ii. end if
        c. end for
    d. Remove the last in { Alive Nodes List }
5. end while
6. generate a TDMA schedule for all the member nodes in every cluster
7. send the TDMA schedule and the details of minimum tree based clustering in network
return {TREE}

```

2 The Sensor Data Fusion

In any WSN, the sensor nodes transmit the sensed information in the form of packets. However, it is not advisable to transmit all the packets received by a node since there may be uncertainties and redundancies in the sensed data. The redundancy in packets leads to unnecessary energy consumption and bandwidth wastage. Also, the sensors may produce some erroneous information due to many reasons like evolving environmental conditions and manufacturing defects. If this erroneous information are transmitted to BS, it will seriously affect the outcome of the decisions made by the WSN. This research work employs Type-2 fuzzy logic to prevent the redundant data from getting transmitted to BS. After the completion of CH selection and MST construction, every CH generates a TDMA schedule and disseminate it to every member of cluster to deliver data. Each live sensor node in the network is associated with a Type-2 Fuzzy Logic Controller (FLC). The FLC finds the confidence factor (CF) of the data sensed by the sensor based on the current sensor condition. Thus each sensor node

generates packets that consist of both the data and the confidence factor of that data. There are three types of sensor input data; Temperature, Humidity, Signal to Noise Ratio (SNR). For each input data, the expected value and its uncertainty are represented by the covariance and mean matrix. The values are then normalized to a value between [0, 1]. The inputs to the fuzzy system are distributed into three levels: Low, Medium, and High. The output of the FLC is the consequent which is broadly divided into five levels: Very Low, Low, Medium, High, and Very High. As there are three states in each input variable (Low, Medium, and High) and there are three variables (Temperature, Humidity, Signal to noise ratio) there is a total possibility of $3 \times 3 \times 3 = 27$ inference rules. The inference rules are shown in Table 1.

Table 1. Inference rules

Rule no.	Input variable			Output
	Temperature	Humidity rate	Signal to noise	Confidence factor
1	High	High	High	Very high
2	High	High	Medium	Very high
3	High	High	Low	High
4	High	Medium	High	High
5	High	Medium	Medium	High
6	High	Medium	Low	Medium
7	High	Low	High	Medium
8	High	Low	Medium	Medium
9	High	Low	Low	Low
10	Medium	High	High	Medium
11	Medium	High	Medium	Low
12	Medium	High	Low	Low
13	Medium	Medium	High	Medium
14	Medium	Medium	Medium	Medium
15	Medium	Medium	Low	Low
16	Medium	Low	High	Medium
17	Medium	Low	Medium	Low
18	Medium	Low	Low	Very low
19	Low	High	High	High
20	Low	High	Medium	Medium
21	Low	High	Low	Low
22	Low	Medium	High	Medium
23	Low	Medium	Medium	Low
24	Low	Medium	Low	Very low
25	Low	Low	High	Low
26	Low	Low	Medium	Very low
27	Low	Low	Low	Very low

The determination of whether the values of the sensor nodes are in the conventional range is performed by the FLC. If the value is in the accepted range the output of FLC is 100%. If the value is out of range then the FLC generates the CF for data collected. The confidence factor is $0\% \leq CF_n \leq 100\%$. Each sensor node compares the confidence factor of the data sensed against a threshold value or cut-off value. This cut-off is set by the users to determine whether the fuzzy amount produced should be measured or not. If the confidence factor of the data is fewer than the cut-off assessment then the data sensed is discarded. Else, the data is transmitted to the parent node. The confidence factor is calculated for the information that has been sensed by the parent node which is used to decide whether to use the information for fusion or to discard it. If the packets need to be discarded then packets received from its child nodes are fused. Otherwise, the data sensed by the parent node is also fused with the data of its children nodes and the fused data is sent to its parent node. The fusion performed by all the non-child nodes is as follows:

$$FD = \frac{(CF_1 \times D_1) + (CF_2 \times D_2) + (CF_3 \times D_3) + \dots + (CF_n \times D_n)}{CF_1 + CF_2 + CF_3 + \dots + CF_n} \quad (8)$$

Where D_1, D_2, \dots, D_n the data are received by the parent node from its child nodes of one kind and CF_1, CF_2, \dots, CF_n are the confidence factors of the corresponding data and FD is the fused data that will be transmitted to its parent node. As data from different nodes are fused together into FD and also that the data that is used for fusion is of high confidence value the FD is robust and is of higher certainty. The FD is calculated independently for each type of sensor nodes and hence we have a set of FDs instead of a single FD. The set of FDs is represented as a vector V_{FD} .

$$V_{FD} = \{FD_1, FD_2, FD_3, \dots, FD_m\} \quad (9)$$

Where m is the number of different types of data being sensed

Suppose a parent node has three child nodes that sense temperature and the three temperatures are 30 °C, 25 °C and 20 °C their corresponding confidence factors are 0.50, 0.75 and 0.65 respectively. Then the FD for the temperature is found to be 25.60. Similarly, the FDs are calculated for other data as well and the vector consisting of FD will be $V_{FD} = \{25.60, 53.2, 38\}$. Then the consequent of the new data is found. This vector is then passed by the parent node to its parent only if the consequent of the new data is changed. But if there is a modification in the arrangement it does not mean a correct exposure. It is instead measured a likely occurrence in the region that is being monitored. The BS regularly processes the received data to determine whether it is an event or not.

Algorithm 3: Data Fusion and Data Transmission

If (node I in {TREE} is parent node) then

- a. If (the data packet need decompressing) then
 - i. For each 10 bits in compressed data packet do
 1. Y =high 7 bits as side information
 2. S_x = lower 3 bits
 3. Calculate $S_y=H*Y^T$
 4. Calculate $X=[S_x \text{ xor } S_y]^T$
 5. Search X in Z_x which has $Z_x \text{ xor } Y = X$
 6. Append Y and X into data packet decompression
 - ii. End For
 - b. End If

//the received data packets contain $D1, D2, D3, \dots, Dn$ data and corresponding confidence factors $CF1, CF2, CF3, \dots, CFn$ //

//The parent node has the input data T -Node Temperature, H -Humidity Ratio, N -Signal to Noise Ratio //
 - c. $CFp = FLC(T, F, N)$
 - d. If $CFp \leq \delta$
 - i. Discard the data collected
 - ii. FDn =Received Data Packet from each child node will be fused
 - iii. Consequent= $FLC(FD1, FD2, FD3 \dots FDn)/n$ is number of child nodes for
Else
 - iv. Generate packet for the collected data
 - v. $FDn+1$ = Received Data Packets from each child node will be fused along with the packet generated for the collected data
 - vi. Consequent= $FLC(FD1, FD2, FD3 \dots FDn, FDn+1)$
 - e. End If
 - f. If the resultant was not altered
 - i. Ignores the conventional data
 - ii. Send only the nodes data
 - g. Else
 - i. Send the Consequent to parent node
 - ii. IF the event sensed by parent node
 1. Report the event
 - iii. End IF
 - h. End If
2. End If
3. If (the node is child node)
- a. $CFc = FLC(T, F, N)$
 - b. If $CFc \geq \delta$
 - i. Data packet data and CFc will be sent to its parent node
 - c. Else
 - i. Discard the data
 - d. End If
4. If (the data packet need compressing) then
- a. For each 14 bits in the data packet do
 - i. Y =high 7 bits as side information
 - ii. X = lower 3 bits
 - iii. Calculate $S_x=H*Y^T$
 - iv. Append Y and S_x^T into new packet
 - b. End For
5. End If

3 Simulation Parameters

To validate the performance of the projected algorithm, the FB-DFTBC, DFTBC, and LEACH-C are implemented using MATLAB where the simulation parameters are given in Table 2.

Table 2. Parameters for simulation

Description of parameter	Value
1. Area of simulation	100 × 100 square meters
2. Number of nodes in network	100
3. E_{elec} (radio electronic energy)	50 nJ/bit
4. E_{fs} (radio free space)	100 pJ/bit/m ²
5. E_{mp} (radio amplifier energy)	0.013 pJ/bit/m ⁴
6. E_{init} (initial energy of node)	2 J
7. Packet size	500 bytes
8. Base station location	50,175
9. Channel type	Wireless channel
10. Number of clusters	5
11. Simulation time	3600 s

4 Evaluation and Results

4.1 The Simulation Results

Firstly, Fig. 4 showcases that the proposed approach enhances the overall network lifetime than their counterparts. Here, the X-axis represents the lifetime of the network in terms of a number of rounds whereas, Y-axis symbolizes the number of nodes alive. From Fig. 4, it is obvious that the FBDF-TBC (blue in color) have more rounds or longer network lifetime compared to LEACH-C and DFTBC, respectively. Besides, Fig. 4 also describes the comparison between the FBDF-TBC with and without compression (red in color). In short, it enhances the overall network lifetime of WSN from a minimum of 16% (compared to FBDF-TBC without DSC) to a maximum of 76% (compared to LEACH). After that, Fig. 5 represents the time of the death of the last node in the network when the location of Base Station is varied.

It is clear from the graph that the number of rounds in DFTBC and FBDF-TBC relatively decreases when the BS is moved farther, but it remained stable with LEACH-C. Subsequently, Fig. 6 quantifies the percentage of dead nodes for five different routing protocols: FB-DFTBC, DFTBC with DSC, DFTBC without DSC, PEGASIS, and LEACH. The X-axis represents the percentage of dead nodes whereas the Y-axis contains the time in terms of a number of rounds at which the particular percentage of nodes are dead. The graph clearly depicts that FB-DFTBC routing protocol takes longer to attain a larger percentage of dead nodes that all other protocols. So it can be stated that it increases the lifetime of the network and the sensor nodes. Packet delay which is defined as the time difference between the time at which the

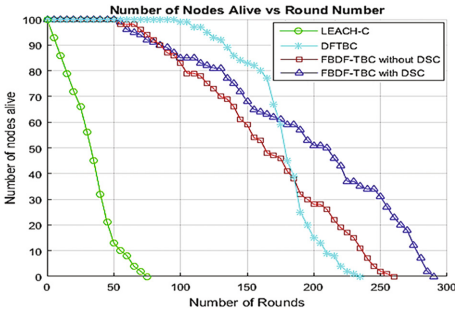


Fig. 4. Number of live nodes with the change in rounds

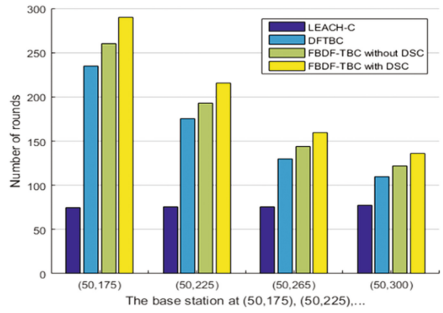


Fig. 5. Death of the last node when the position of BS changes

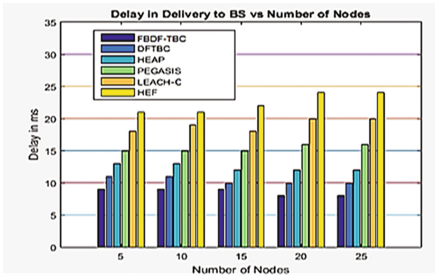


Fig. 7. Delay in the delivery of packets to BS

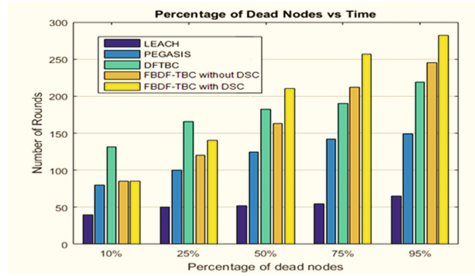


Fig. 6. Percentage of dead nodes when BS at 50,175

packet is created and the time at which it is actually received by the BS is depicted in Fig. 7. It showcases the packet delay of the following WSN routing protocols: FB-DFTBC, DFTBC, HEAP, PEGASIS, LEACH and HEX. Here, the X-axis is the number of nodes that are alive in the sensing environment while the Y-axis is the delay in delivery to BS in milliseconds (ms). It is clear from the graph that the delay is the minimum for FB-DFTBC.

5 Conclusion and Future Enhancements

Generally, energy preservation is the major focus in any wireless sensor network research. With the similar objective, this research also work proposes a new approach for energy-efficient clustering and compression of data packets before actually sending it to the base station. Subsequently, it is also proved from the experimental results that the proposed algorithm greatly minimizes the energy consumption of the network which in turn improves the overall network lifetime. Further, it also enhances the sensing accuracy by eliminating the data redundancy. In concise, the simulation results clearly

concludes that the energy efficiency of the proposed algorithm is higher than its peers thereby improving the network lifetime from a minimum of 16% to a maximum of 76%. The proposed approach can be further enhanced by incorporating sub-clustering as well as fault tolerant characteristics which are our ongoing research work.

References

1. Kumar, N., Tyagi, S., Deng, D.: LA-EEHSC: learning automata-based energy efficient heterogeneous selective clustering for wireless sensor networks. *J. Netw. Comput. Appl.* **46**, 264–279 (2014)
2. Yu, J., Feng, L., Jia, L., Gu, X., Yu, D.: A local energy consumption prediction-based clustering protocol for wireless sensor networks. *Sensors (Switzerland)* **12**, 23017–23040 (2014)
3. Tan, N.D., Viet, N.D.: DFTBC: Data Fusion And Tree-Based Clustering Routing Protocol for Energy-Efficient in Wireless Sensor Networks, vol. 326, pp. 61–77. Springer, Cham (2015)
4. Sengaliappan, M., Marimuthu, A.: Improved general self-organized tree-based routing protocol for wireless sensor network. *J. Theor. Appl. Inf. Technol.* **1**, 100–107 (2014)
5. Liu, Z., Yang, X.: An application model of fuzzy clustering analysis and decision tree algorithms in building web mining. *Int. J. Digit. Content Technol. Appl.* **23**, 492–500 (2012)
6. Mammu, A., Hernandez, S.K., Jayo, U., Sainz, N., de la Iglesia, I.: Cross-layer cluster-based energy-efficient protocol for wireless sensor networks. *Sensors (Switzerland)* **4**, 8314–8336 (2015)
7. Su, S., Yu, H., Wu, Z.: An efficient multi-objective evolutionary algorithm for energy-aware QoS routing in wireless sensor network. *Int. J. Sens. Netw.* **13**(4), 208–218 (2013)
8. Izadi, D., Abawajy, J.H., Ghanavati, S., Herawan, T.: A data fusion method in wireless sensor networks. *Sensors (Switzerland)* **2**, 2964–2979 (2015)
9. Zhang, Z., Liu, T., Zhang, W.: Novel paradigm for constructing masses in Dempster-Shafer evidence theory for wireless sensor network's multisource data fusion. *Sensors (Switzerland)* **4**, 7049–7065 (2014)
10. Chen, J., Han, X.: The distributed source coding method research based on clustering wireless sensor networks. *Int. J. Sens. Netw.* **4**, 224–228 (2015)
11. Bagci, H., Yazici, A.: An energy aware fuzzy approach to unequal clustering in wireless sensor networks. *Appl. Soft Comput. J.* **4**, 1741–1749 (2013)
12. Meghanathan, N.: Stability-based and energy-efficient distributed data gathering algorithms for wireless mobile sensor networks. *Ad Hoc Netw.* **19**, 111–131 (2014)
13. Heinzelman, W., Chandrakasan, A., Balakrishnan, H.: An application specific protocol architecture for wireless microsensor networks. *Proc. IEEE Wirel. Netw.* **4**, 660–670 (2002)
14. Gupta, I., Riordan, D., Sampalli, S.: Cluster-head election using fuzzy logic for wireless sensor networks. In: *Annual Conference on Communication Networks Services*, pp. 255–260 (2005)
15. Aderohunmu, F., Deng, J., Purvis, M.: A deterministic energy-efficient clustering protocol for wireless sensor networks. In: *Proceeding of the Seventh International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)*, pp. 341–346 (2011)
16. Natarajan, H., Selvaraj, S.: A fuzzy based predictive cluster head selection scheme for wireless sensor networks. In: *International Conference on Sensing Technology*, pp. 560–566 (2014)

Biologically-Inspired Foraging Decision Making in Distributed Cognitive Radio Networks

Olukayode A. Oki¹(✉), Thomas O. Olwal², Pragasen Mudali¹,
and Matthew Adigun¹

¹ University of Zululand, KwaDlangezwa, X1001, Richards Bay 3886,
Republic of South Africa
okikayode@gmail.com

² Tshwane University of Technology, Pretoria 0001, Republic of South Africa

Abstract. The dynamic spectrum management techniques have been introduced to address the current Radio Frequency bands inefficiency challenges. Cognitive Radio (CR) technology has been regarded as the most promising technology in the dynamic spectrum management area. One of the major aspects of the spectrum management is the decision making ability of CR users. The dynamic reconfiguration of both the operating frequency and channel bandwidth in a distributed CR network has not received sufficient attention despite their importance in spectrum decision making. Few research works have attempted to address the dynamic reconfiguration of frequency and channel bandwidth problems using various approaches. However, due to certain challenges such as high computational complexity, ambiguity, repeatability and the lack of optimality with the existing approaches, researchers are still trying to explore newer methods that can achieve optimal spectrum management. Hence, this paper presents a biologically-inspired optimal foraging model for dynamic reconfiguration of frequency and channel bandwidth in a distributed cognitive mobile adhoc network. One of the main advantages of biologically-inspired foraging model is its analytical simplicity and optimum solution. The mean efficiency and Distance travelled by SUs before finding available frequency were measured. The two metrics were measured when subjected to different SUs positions and Giving-Up Time. It was generally observed that the SUs perform better when $0 < X_o \leq 0.2$ and $GUT \leq 50$ in the achieved mean efficiency and distance travelled to find available frequency.

Keywords: Cognitive radio · Distributed · Decision making · Foraging · Spectrum

1 Introduction

In recent years, the remarkable rapid evolution of wireless communication technologies and smart devices that enable social networking applications and other multimedia-based services, has led to the limitation of radio frequency spectrum which is fast becoming scarce. The limited available frequency bands and the inefficient usage of the

Radio Frequency (RF) bands have necessitated the need for a new or a better way of assigning RF spectrum. Dynamic spectrum access and management techniques have been introduced to address the current RF bands inefficiency challenges. Cognitive Radio (CR) technology is being regarded as the most promising technology in the dynamic spectrum management area. CR technology promises to address the depletion and inefficient utilization of spectrum by opportunistically accessing the usable spectrum in an optimal manner. A Cognitive Radio network (CRN) is an intelligent wireless transmission system that possesses the ability to change its transceiver parameters such as frequency and channel bandwidth based on the interaction (e.g. spectrum sensing) with the environment wherein it operates [1, 2].

The basic idea of a CRN is that it should be capable of sharing available frequency bands amongst the licensed/Primary Users (PUs) and unlicensed/Secondary Users (SUs). However, the CRNs operate under the bounded constraint that the PU transmissions should not be interfered with [3] by SUs. Hence, as soon as PU activities are detected on a given channel, the SU must immediately vacate the channel and continue its transmission on another available channel.

In order to realize an efficient utilization of spectrum in a CR environment, a dynamic framework for spectrum management is required. This dynamic spectrum management comprises: spectrum sensing, decision making, sharing and spectrum mobility. The ability of the SUs to select the best accessible spectrum band to fulfil users Quality of Service (QoS) requirements is termed as spectrum decision making; which comprises of three major functions; spectrum characterization, spectrum selection and dynamic reconfiguration of cognitive radio [4].

As with traditional wireless networks, a CRN topology can be classified as either a centralized (infrastructure-based) or a distributed (infrastructure-less or ad-hoc based) network topology. In the centralized topology, a central node such as a base station or access point is deployed with several SUs associated with it. SUs communicate directly with each other, without a central or controlling node, in the distributed topology.

Centralized and adhoc networks are usually characterized by a fixed and low number of supported channels (mostly less than ten or at most in order of tens). However, spectrum decision-making in a distributed CRN, where the number of supported channel ranges in the order of thousands, is a serious challenge that needs to be addressed [5–7].

In a distributed CRN environment such as in a mobile adhoc network, when there are several frequencies and channel bandwidths available, dynamically selecting the best combination of frequencies and bandwidths is an important challenge. The complexity of this challenge is increased when spectrum quality and the QoS requirements of various application types are considered. The diversity of spectrum bands and the guiding principles issued by the communications regulatory agencies for how to access the spectrum implies that the CR nodes for mobile adhoc networks should dynamically reconfigure their operating frequency and channel bandwidths, as network conditions dictate.

The dynamic reconfiguration of both the operating frequency and channel bandwidth in a distributed CR network has not received sufficient scholarly attention despite its importance in spectrum decision making [5, 7]. Various research works have attempted to solve the decision making problem using various approaches, such as

theoretical, statistical, predictive CCC, etc. However, these approaches suffer from challenges such as high computational complexity, ambiguity, repeatability and applicability. Thus, researchers are still trying to explore other approaches that can be used to address the challenges with existing approaches and to achieve optimal spectrum management. Our previous study [8] has already discussed the existing approaches and their challenges. The study subsequently introduced a biologically-inspired foraging approach to address the decision making problem and other problems relating to the existing approaches. The biologically-inspired foraging approach has been described and is being adopted by many researchers in the field of communication networks due to its analytical simplicity and its generic applications.

The results from previous studies [16–18] in other wireless networks, shows that the biologically-inspired foraging approach has generic, simple and high applicability properties. In studies [16, 17], the biologically-inspired foraging approach was utilised to develop the BEACH and FIRE-MAN protocols respectively. Both studies evaluated the developed protocols performances by measuring both the throughput and energy-efficiency in distributed heterogeneous networks. Based on the results of their studies, it was observed that the proposed biologically-inspired foraging approach protocols performs better than other conventional approaches in the field of heterogeneous networks.

Hence, this paper presents a biologically-inspired, optimal composite foraging model in addressing the dynamic reconfiguration of frequency and channel bandwidth in distributed CRNs. One of the main advantages of the biologically-inspired foraging model is its analytical simplicity and optimum solution. This advantage will help to address the shortcomings with other existing approaches and will also help to achieve optimum spectrum management.

To the best of our knowledge, this work can be viewed as an early contribution towards the application of the composite foraging theory of Nutrients Optimisation to the field of distributed CRN research.

The remainder of this paper is arranged as follows: Section 2 presents Biologically-inspired foraging theory. Section 3 proposes the biologically-inspired composite foraging algorithm for CRN decision making. Section 4 presents the model analytical solution and discusses the results obtained. The paper is concluded in Section 5 with an outline of the future work.

2 The Biologically-Inspired Optimal Foraging Theory

The study of how natural foraging animals in an arbitrary environment make optimal decisions is referred to as biologically-inspired optimal foraging theory. The optimal decisions made by foragers help them to maximize their efficiency, have long lifetime and reduces possible threats. Foraging theory uses diverse models to describe how solitary foraging animals search for prey sorts and make optimal decision on which prey to feed on, so as to maximize their efficiency [9]. The classifications of nutrients consumption by foraging animals was adopted and modeled as an optimization process, which is now commonly known as optimal foraging theory. The ability of a forager to make an optimal decision on the most suitable prey type to consume so as to maximize

their efficiency within the smallest possible time interval, is one of the core advantages of the optimal foraging theory. In spite of this advantage, one of the major factors that influence the foraging efficiency is the selection criteria used by the forager in selecting a prey type to consume. Hence, foragers should aim to match their search effort to the relative profitability of various parts of their surroundings.

One of the application area of optimal foraging theory is in the decision making of SUs in cognitive radio network. Optimal foraging theory can be used to model the decision making for SUs in selecting an appropriate frequency and channel bandwidth for communication as illustrated in Fig. 1.

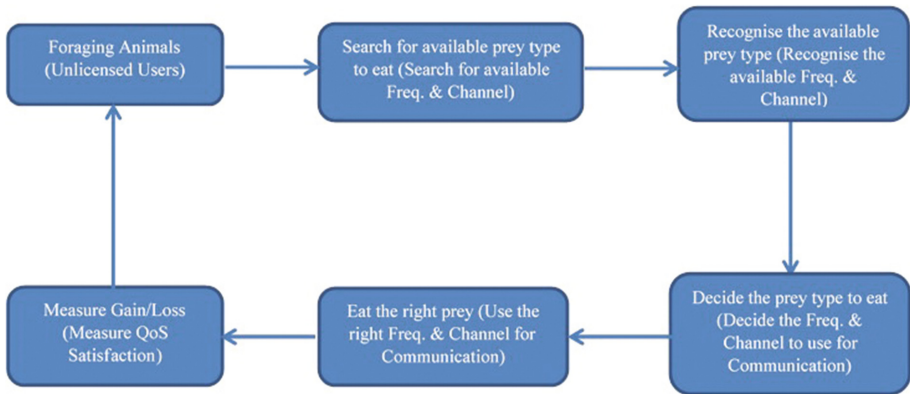


Fig. 1. Optimal foraging cycle for distributed CRNs [8]

There are many existing biologically-inspired optimal foraging models [10–13]. However, one of the most common among these models is the composite prey model. Apart from being one of the most common among the optimal foraging models, this study adopts the composite model also because of its effectiveness and applicability to the decision making for a distributed cognitive mobile adhoc network.

The composite model assumes that there are “n” different types of prey, which the foragers can consume for energy. P_i is the relative frequency or the probability of a forager encountering a prey type i . The average rate of encountering a prey type i is λ_i and V_i is the expected amount of energy intake from captured prey i . While T_i is the expected time to seek and capture prey i . Hence, the efficiency E of a forager can be defined as the ratio of the expected intake energy to the time spent by the forager. This is represented mathematically as:

$$E = \frac{\sum_{i=1}^n P_i \lambda_i V_i}{\sum_{i=1}^n P_i \lambda_i T_i} \tag{1}$$

The maximization of the foragers efficiency, E , involves finding the optimal values of P_i for all prey i . Based on the zero-one rule used by foragers to determine optimal values of P_i , it can be noticed that this theory establishes a solid basis for decision making approach and optimization problems.

3 The Biologically-Inspired Optimal Foraging Theory

This section presents the analogy between the biologically-inspired foraging composite model and distributed CRNs. To create a biologically-inspired foraging communication for CRNs, we considered the Secondary Users (SUs) to be the foragers, and prey to be the available Primary Users (PUs) frequency. In analogous to the biological forager, whereby the forager searches for prey just the same way each SU node with a message searches for possible available PUs frequency to be used for communication. Hence, in this model, each of the SUs node selects and uses available PUs frequency for communication, in order to maximize the spectrum utilisation with minimum interference to the PU network.

In biological composite search technique, the foragers can use two stages (Intensive and/or Extensive) to search across the search space [14]. The first stage of the search involves an intensive search, which is characterised by frequent changes of direction effected by making small steps in the search area. Hence, this search mode is usually area-intensive. The kind of motion described in the intensive search mode is based on the Brownian (non-heavy-tailed) motion. However, if this technique has not been successful, i.e. if no frequency has been encountered by SUs after a particular time σ , known as the Giving-Up Time (GUT), the SU switches to the extensive search mode by taking relatively longer steps, using ballistic motion, with lesser change of direction [See Fig. 2].

The overall objective is to find an optimal value of GUT (σ) that minimizes the expected distance travelled by an SU before finding an available frequency. In order to achieve this, the following simplifying assumptions have to be made:

1. The SUs search for available frequency in a one-dimensional line
2. The SU starts at a position X_0 (where it last found a food item)
3. The SU uses a Brownian movement pattern in the intensive mode and a ballistic movement in the extensive mode
4. The distance the SUs must travel before finding a food item after switching to the extensive mode is exponentially distributed with mean $D = \frac{d}{2}$. This distance is independent of the position of the SUs, i.e. $X(\sigma)$, after completing the intensive search.

In biological foraging environment, the foragers usually start to search for prey within their immediate environment, and only search long distance, if no prey was found. Hence, this study starts with the Brownian type random search movement in which the forager moves in the intensive mode, described by the stochastic differential equation:

$$dX(t) = \alpha dW(t) \quad (2)$$

where $X_{(t)}$ is the position of the SUs at any time t , $W_{(t)}$ is a Wiener process with parameter τ^2 (variance). Now, the mean instantaneous speed of the intensive search process described by Eq. (2) is:

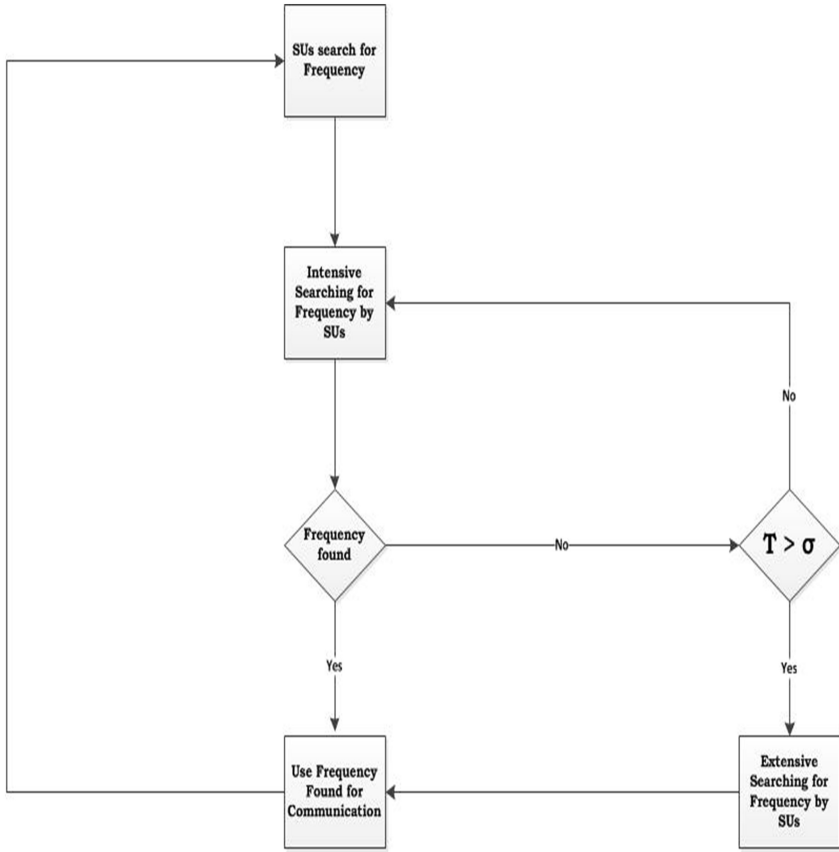


Fig. 2. Composite search flow process

$$v_I = \sqrt{\frac{2}{\pi\tau}} \quad (3)$$

Thus, the total distance travelled during the intensive phase is:

$$d = v_I T \quad (4)$$

At $T > \sigma$ (Expected time to seek for available frequency $>$ GUT), the SUs switches to the extensive search mode. The distance travelled by the SUs between successive available frequency, L , is given for the two search phases as:

$$L = \begin{cases} v_I T & \text{if } T \leq \sigma(\text{Intensive}); \\ v_I T + R & \text{if } T > \sigma(\text{Extensive}). \end{cases} \quad (5)$$

Here, random variable R is a distance taken from an exponential distribution, i.e. $R \sim e^{\frac{1}{\sigma}}$.

The probability density function of the total distance travelled by SUs (from Eq. (5)) before finding available frequency is given as:

$$f(l) = \begin{cases} \frac{1}{v_I} f_T\left(\frac{l}{v_I}\right) & \text{for } l \leq v_I \sigma; \\ (1 - F_T(\sigma)) f_R(l - v_I T) & \text{for } l > v_I \sigma. \end{cases} \quad (6)$$

It is assumed in this work that the energy cost during search by the SUs is directly proportional to distance travelled, and that the energy obtained from each available frequency found is the same. Hence, by minimizing the mean distance travelled between available frequencies, the SUs will be able to maximize their net energy gain.

From Eq. (6), the expected distance travelled $E(L)$ by the SUs before finding an available frequency for communication is:

$$E(L) = \int_0^{\infty} l f_L(l) dl = v_I \int_0^{\infty} t f_T(t) dt + (1 - F_T(\sigma)) \times \left(v_I \sigma + \int_0^{\infty} s f_R(s) ds \right). \quad (7)$$

The Eq. (7) can be written in a closed form as:

$$\frac{E(L)}{d} = \sqrt{\frac{4x_0^2 \sigma}{\pi d^2}} e^{\left(-\frac{x_0}{\pi v_I \sigma}\right)} + \left(\frac{2x_0^2}{\pi v_I d} + \frac{v_I \sigma}{d} + \frac{1}{2} \right) \operatorname{erf} \left(\sqrt{\frac{x_0^2}{\pi v_I^2 \sigma}} \right) - \frac{2x_0^2}{\pi v_I d}. \quad (8)$$

If $\varphi = \frac{v_I \sigma}{d}$ and $\in = \frac{2x_0^2}{\pi v_I d}$, then,

$$\frac{E(L)}{d} = \sqrt{\frac{2 \in \varphi}{d}} e^{\frac{\in}{2\varphi}} + \left(\in + \varphi + \frac{1}{2} \right) \operatorname{erf} \left(\sqrt{\frac{\in}{2\varphi}} \right) - \in. \quad (9)$$

Here, erf means error function. Suppose there is a local minimum in $E(L)$ for a positive value of φ , then this will occur where the derivative of $E(L)$ with respect to φ is 0. That is:

$$\frac{dE(L)}{d\varphi} = 0$$

This corresponds to:

$$\frac{\in^{\frac{1}{2}}}{2^{\frac{3}{2}} \pi^{\frac{1}{2}} \varphi^{\frac{3}{2}}} e^{-\left(\frac{\in}{2\varphi}\right)} - \operatorname{erf} \left(\sqrt{\frac{\in}{2\varphi}} \right) = 0 \quad (10)$$

From Eq. (10), there is no closed form solution to obtain the optimal value of φ^* , since (10) is a transcendental equation for φ . Hence, the only way forward is to consider the two terms in Eq. (10) separately. That is, let

$$A(\varphi) = \frac{\epsilon^{\frac{1}{2}}}{2^{\frac{3}{2}}\pi^{\frac{1}{2}}\varphi^{\frac{3}{2}}} e^{-\left(\frac{\epsilon}{2\varphi}\right)} \text{ and } B(\varphi) = \operatorname{erf}\left(\sqrt{\frac{\epsilon}{2\varphi}}\right)$$

So that Eq. (10) becomes:

$$A(\varphi) - B(\varphi) = 0. \quad (11)$$

Therefore, two cases arise:

CASE 1: If $A(\varphi) < B(\varphi)$ for all $\varphi \geq 0$, then, $E(L)$ is a monotone increasing function of φ . While the optimal value of φ , i.e. φ^* , occurs at $\varphi^* = 0$.

CASE 2: If $A(\varphi) > B(\varphi)$, $E(L)$ becomes a decreasing function of φ . In this case, a local maximum and minimum exist in $E(L)$. The local minimum turns out to be a global minimum if it is smaller than the value of $E(L)$ at $\varphi = 0$.

Suppose the local maximum occurs at $\varphi = \varphi_a$. Thus, by solving $A(\varphi_a) = 0$, one obtains $\varphi_a = \frac{\epsilon}{3}$. The implication of this is that $A(\varphi) > B(\varphi)$ if and only if

$$\frac{3^{\frac{3}{2}}}{2^{\frac{3}{2}}\pi^{\frac{1}{2}}e^{\frac{3}{2}}\epsilon} > \operatorname{erf}\left(\frac{3^{\frac{1}{2}}}{2^{\frac{1}{2}}}\right) \quad (12)$$

This means,

$$\epsilon_x < \epsilon \simeq \frac{1}{4}$$

To obtain an approximation to Eq. (10), we assume that $K = \frac{1}{\varphi}$ and then find the power series representations of A and B in K . Hence, we have:

$$\frac{2^{\frac{1}{2}}}{\pi^{\frac{1}{2}}}\left(\left(\epsilon k\right)^{\frac{1}{2}} - \frac{\left(\epsilon k\right)^{\frac{3}{2}}}{6} + \frac{\left(\epsilon k\right)^{\frac{5}{2}}}{40} + \mathcal{O}\left(\left(\epsilon k\right)^{\frac{7}{2}}\right)\right) - \frac{\epsilon^{\frac{1}{2}}}{2^{\frac{3}{2}}\pi^{\frac{1}{2}}}k^{\frac{3}{2}}\left(1 - \frac{\epsilon k}{2} + \frac{\left(\epsilon k\right)^2}{8} + \mathcal{O}\left(\left(\epsilon k\right)^3\right)\right) = 0$$

Collecting the powers of K gives:

$$1 - \left(\frac{\epsilon}{6} + \frac{1}{4}\right)k + \epsilon\left(\frac{\epsilon}{40} + \frac{1}{8}\right)k^2 - \frac{\epsilon^2}{24}k^3 + o\left(\left(\epsilon k\right)^3\right) = 0 \quad (13)$$

Recall that in Case II, we have a solution to Eq. (10), that is $\epsilon \ll 1$ by Eq. (12). Thus, we may seek a solution as power series in ϵ as follows:

$$k = \sum_{n=0}^{\infty} k_n \epsilon^n \quad (14)$$

Substituting (14) into (13) and equating coefficients of ϵ^n ($n = 0, 1, 2$) give the coefficients in the series for K as follows:

$$k_0 = 4, \quad k_1 = \frac{16}{3}, \quad k_2 = \frac{392}{45}$$

Suppose (12) is satisfied, then there exists a local minimum $E(L)$ with respect to σ and this occurs at:

$$\begin{aligned} \sigma^* &= \frac{d}{v_I} \left(4 + \frac{16}{3} \epsilon + \frac{392}{45} \epsilon^2 + O(\epsilon^3) \right)^{-1} \\ &= \frac{d}{4v_I} \left(1 - \frac{4}{3} \epsilon - \frac{2}{5} \epsilon^2 + O(\epsilon^3) \right) \end{aligned} \quad (15)$$

Disregarding the terms of order ϵ^3 in the Eq. (15) above gives:

$$= \frac{d}{4v_I} \left(1 - \frac{4}{3} \epsilon - \frac{2}{5} \epsilon^2 \right) \quad (16)$$

The optimal value of σ is the one that minimizes $E(L)$ and can be approximated by Eq. (16). If $\epsilon > 0.2$, then the minimum value of $E(L)$ occurs at $\sigma = 0$. From Eqs. (16) and (9), it can be deduced that if frequency are densely distributed relative to the number of SUs within an environment, $\epsilon > 0.2$, then the optimal strategy is always to search using straight line (extensive), until a frequency is found. And the mean distance travelled is $d/2$. However, if $\epsilon < 0.2$, then the mean efficiency is improved by using Brownian motion (Intensive) to search. The optimal duration of intensive searching increases as ϵ decreases.

4 The Analytical Solution

This section presents the analytical solution for the model derived above. The mean efficiency and distance travelled by SUs before getting a frequency for communication when subjected to different SUs node position (X_o) and GUT (σ) were analytically simulated. The mean efficiency is the reciprocal of distance travelled by the SUs before finding an available frequency for communication ($1/E(L)$).

Figure 3 shows the mean efficiency over a range of starting positions for different SUs. The mean efficiency at the optimal switching time were calculated using Eqs. (8) and (16). It can be observed that each of the SUs considered, have various efficiency value, however, from $X_o = 0.2$, the mean efficiency of each of the SUs remain constant. This behaviour is in line with the mean exponential distribution $\frac{d}{2}$, which is the efficiency of a ballistic strategy in a random environment. It can also be observed that while $0 < X_o \leq 0.2$, the smaller the X_o , the higher the mean efficiency.

The effects of different SUs positions on the distance travelled before finding an available frequency for communication were presented in Fig. 4. The distance travelled by SUs were calculated using Eq. (8). It was observed that as X_o increases, the distance travelled by SUs ($E(L)$) tends towards $R \sim e^{\frac{2}{d}}$, exponential distribution.

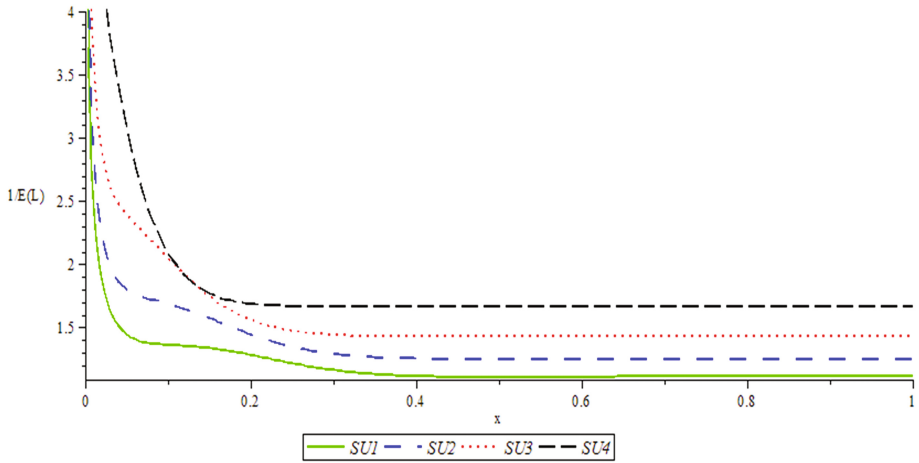


Fig. 3. The effect of SUs positions on mean efficiency

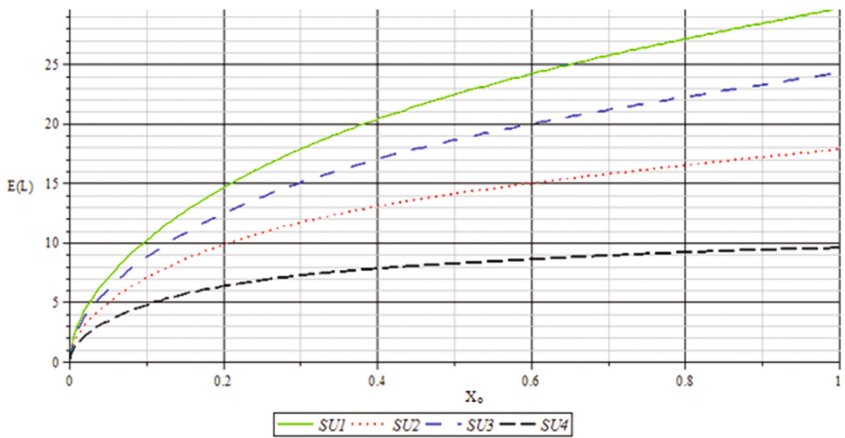


Fig. 4. The effects of SUs positions on distance travelled before getting available frequency

Figures 5 and 6 present the SUs mean efficiency, when subjected to different GUT (σ). Two sets of GUT were considered; early GUT ($10 \leq \sigma \leq 50$) and late GUT ($60 \leq \sigma \leq 300$). The early and late GUT values were calculated by numerical minimization of Eq. (7) and using the approximation of Eq. (16). The reciprocal of Eq. (8) were used to calculate the mean efficiency, which is plotted against various GUT. The optimal value of σ is the one that minimizes $E(L)$ and can be approximated by Eq. (16). It can be observed from Fig. 5 that the mean efficiency drops significantly as σ increases. However, when $\sigma > 50$, it was observed that the mean efficiency drops drastically to almost zero, as depicted in Fig. 6. The very low mean efficiency achieved could be attributed to the delay incurred when $\sigma > 50$. The delay could lead to the

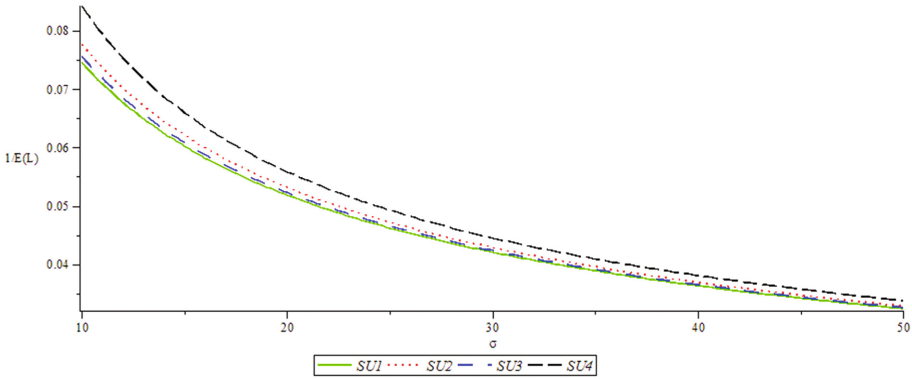


Fig. 5. The effect of early GUT on mean efficiency

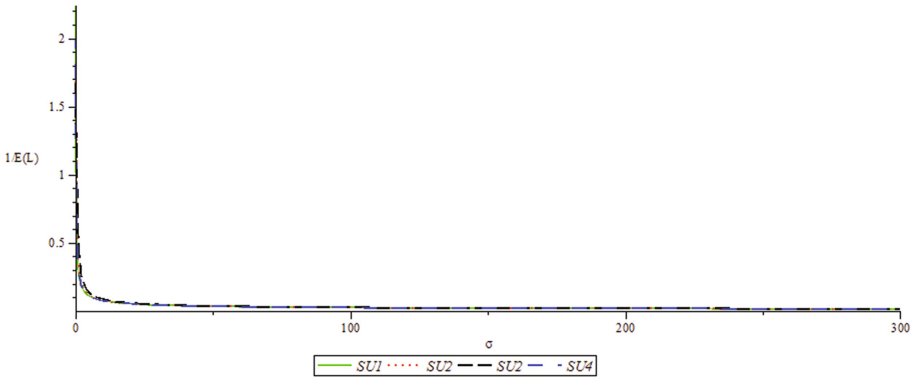


Fig. 6. The effects of late GUT on mean efficiency

arrival of PUs, hence, it makes the frequency not available again for SUs communication, which in turn would lead to low mean efficiency.

Figures 7 and 8 show the Distance travelled by the SUs before finding available frequency for communication ($E(L)$) against early and late GUT (σ) respectively. The $E(L)$ were calculated using Eq. (7) and the lower the $E(L)$ the better the performance. It can be observed from Figs. 7 and 8 that as σ increases, the $E(L)$ also increases exponentially. Based on our observation, the model performs better while $\sigma \leq 50$. It is generally assumed that the higher the distance travelled, the higher the energy expenditure and that in biological foraging, the energy obtained from each food item found are the same [15]. Hence, by minimizing the distance travelled before finding available frequency, the SUs will be able to maximize its overall energy gain and achieve high efficiency.

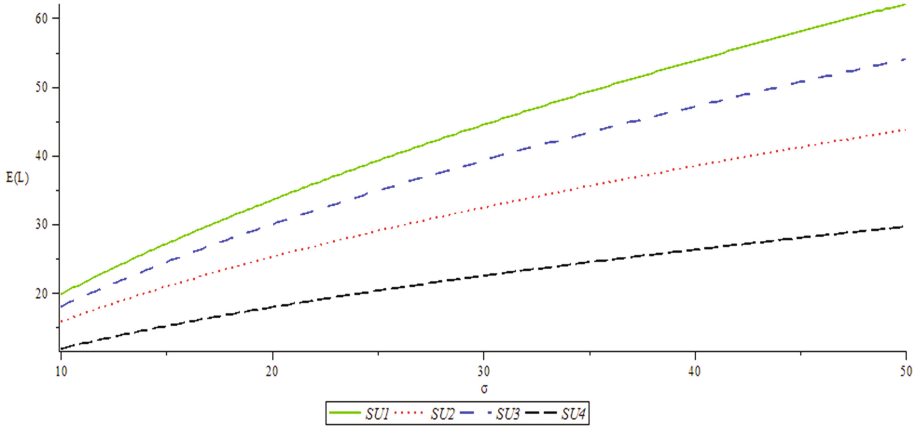


Fig. 7. The effect of early GUT on distance travelled by SUs before finding available frequency

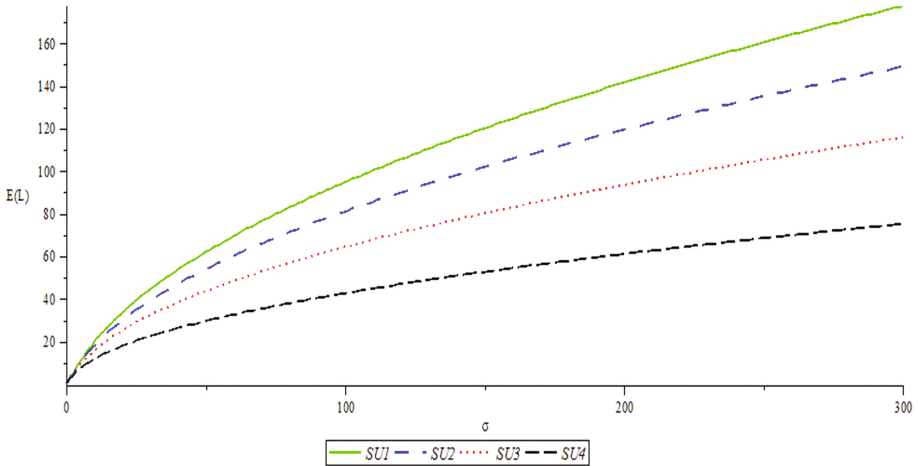


Fig. 8. The effect of late GUT on distance travelled by SUs before finding available frequency

5 Conclusion

In conclusion, this paper has presented a biologically-inspired foraging theory for decision making in distributed Cognitive radio networks. The composite model in foraging theory was used and it comprises both intensive and extensive random search model. The intensive search model used Brownian movement pattern while the ballistic movement pattern was used for the extensive movement. The SUs start the search for available frequency from intensive mode. However, if no frequency is encountered after specified period called GUT, the SUs switch to the extensive search mode.

The mean efficiency and Distance travelled by SUs before finding available frequency when subjected to different SUs positions and GUT were measured. It was

generally observed that the SUs perform better when $0 < X_o \leq 0.2$ and $\sigma \leq 50$, in the achieved mean efficiency and distance travelled to find available frequency. The analytical solution also shows that as ϵ increases and the SUs moves away from the patch, the GUT predicted by intensive and extensive distances also increases, however, the time predicted by minimizing $E(L)$ decreases.

In future, we intend to implement and validate this model, using computer simulations and to measure energy efficiency, successful transmission probability and average throughput of the model presented.

References

1. Masonta, M.T., Mzyece, M., Ntlatlapa, N.: Spectrum decision in cognitive radio networks: a survey. *IEEE Commun. Surv. Tutor.* **15**(3), 1088–1107 (2013)
2. Mitola, J.: Cognitive radio – model-based competence for software radios. Licentiate thesis, KTH, Stockholm, September 1999.
3. Farzad, H., Sumit, R.: Capacity considerations for secondary networks in TV white space. *IEEE Trans. Mobile Comput.* 1–29 (2013). arXiv: 1304. 1785v1
4. Marinho, J., Monteiro, E.: Cognitive radio: survey on communication protocols, spectrum decision issues and future research directions. *J. Wirel. Netw.* **18**(2), 147–164 (2012)
5. Dere, B.A., Bhujade, S.: An efficient spectrum decision making framework for cognitive radio networks. *Int. J. Innov. Sci. Modern Eng. (IJISME)* **3**(2), 45–48 (2015)
6. Akyildiz, I.F., Won-Yeol, L., Vuran, M.C., Mohanty, S.: A survey on spectrum management in cognitive radio networks. *IEEE Commun. Mag.* **2**(3), 40–48 (2008)
7. Sengupta, S., Subbalakshmi, K.P.: Open research issues in multi-hop cognitive radio networks. *IEEE Commun. Mag.* **2**(3), 168–176 (2013)
8. Oki, O.A., Olwal, T.O., Mudali, P., Adigun, M.O.: Dynamic spectrum reconfiguration for distributed cognitive radio networks. *J. Intell. Fuzzy Syst.* **32**(4), 3103–3110 (2017)
9. Atakan, B., Akan, O.B.: Biological foraging-inspired communication in intermittently connected mobile cognitive radio ad hoc networks. *IEEE Trans. Veh. Technol.* **61**(6), 2651–2658 (2013)
10. Passino, K.M.: Biomimicry of bacterial foraging for distributed optimization. *IEEE Control Syst. Mag.* **22**(3), 52–67 (2002)
11. Quijano, N., Passino, K.M., Andrews, B.W.: Foraging theory for multi-zone temperature control. *IEEE Comput. Intell. Mag.* **1**(4), 18–27 (2006)
12. Stephen, D., Krebs, J.: *Foraging Theory*. Princeton University Press, Princeton, NJ (1986)
13. Olwal, T.O., Djouani, K., Kurien, A.M.: A survey of resource management toward 5G radio access networks. *IEEE Commun. Surv. Tutor.* **18**(3), 1656–1686 (2016). (Third Quarter)
14. Plank, M.J., James, A.: Optimal foraging: Levy pattern or process. *J. R. Soc. Interface* **5**(26), 1077–1086 (2008)
15. Nolting, B.C.: Random search models of foraging behaviour: theory, simulation and observation. PhD thesis, University of Nebraska, Nebraska (2013)
16. Olwal, T.O., Masonta, M.T., Mekuria, F.: Bio-inspired energy and channel management in distributed wireless multi-radio networks (BEACH). *IET Sci. Meas. Technol.* **8**(6), 380–390 (2014)
17. Olwal, T.O., Van Wyk, B.J., Kogeda, O.P., Mekuria, F.: FIREMAN: foraging-inspired radio-communication energy management for green multi-radio networks. In: *Green Networking and Communications*, pp. 29–46. CRC Press, New York (2013)
18. Yu, R.F., Huang, M., Tang, H.: Biologically inspired consensus-based spectrum sensing in mobile ad hoc networks with cognitive radios. *IEEE Netw. Mag.* **2**(3), 26–30 (2011)

Efficient Algorithms for Hotspot Problem in Wireless Sensor Networks: Gravitational Search Algorithm

Srikanth Jannu¹(✉), Suresh Dara³, Katha Kishor Kumar⁴,
and Sabitha Bandari²

¹ Department of Computer Science and Engineering, Vagdhevi Engineering College,
Bollikunta, Warangal 506005, Telangana, India

j.srikanth@live.com

² Department of Electronics and Communications Engineering,
Vagdhevi Engineering College, Bollikunta, Warangal 506005, Telangana, India

sabithabandari@hotmail.com

³ Department of Computer Science and Engineering,
B.V. Raju Institute of Technology (UGC Autonomous),
Narsapur 502313, Telangana, India

darasuresh@live.in

⁴ Department of Computer Science and Engineering, Kakatiya University,
Warangal 506009, Telangana, India

k_kishorkumar@yahoo.com

Abstract. Energy conservation of sensor nodes (SNs) is the major concern of wireless sensor networks (WSNs) as those are operated by small batteries with a limited power. In a clustered WSN, cluster heads (CHs) collect local information such as temperature, humidity, pressure etc. from the member SNs, aggregate it and send to the sink through few intermediate CHs. Here, the CHs that are closer to the sink are overburdened as they are responsible for forwarding more number of packets than the farther CHs that tend to exhaust their energy quickly. This results in network partitioning and this problem is well known as hot spot or energy hole problem. In this paper, a Gravitational Search Algorithm (GSA) approach based clustering and routing algorithms are proposed to address the hot spot problem. In clustering, we select few efficient SNs as CHs from the normal SNs with respect to certain cost function. We design an algorithm for CH selection based on GSA and assign the remaining SNs to the CHs based on another derived cost function. Then, a GSA based routing algorithm is presented with respect to the routing cost function. These algorithms are intended to develop to enhance the lifetime of network with efficient encoding schemes of GSA. The proposed algorithms are simulated on various scenarios of WSNs by varying number of SNs. The results of the proposed algorithms are compared with few well known algorithms to show the supremacy in terms of network lifetime, residual energy and number of alive SNs.

1 Introduction

In wireless sensor networks (WSNs), clustering and routing are two efficient techniques that have been studied extensively for conserving energy [1]. In a clustered WSN, cluster heads (CHs) forward the data packet through the intermediate CHs to the sink. Here, the CHs closer to the sink are over burdened with forwarding packets of farther CHs along with its own packets that tends to pre-maturely exhaust. As a result, some area that is closer to the sink may remain uncovered due to the dead sensor nodes (SNs)/CHs and this phenomena is called as energy hole or hot spot problem. As a solution of this problem, at first, we design clustering algorithm then routing algorithm is designed by using gravitational search algorithm (GSA). In large scale WSNs, the selection of m CHs among n SNs produces *Mycombk* solutions that is known to be an NP-hard problem. Similarly, the routing is also well known as NP-hard problem [7]. On the other hand, the meta-heuristic algorithms are proved that those are suitable for solving such kind of NP-Hard problems. The meta-heuristic algorithm effectively used to solve WSNs such as clustering and routing solved by PSO [3], unequal clustering and routing by CRO [11] and fault-tolerant problem by PSO [4]. Therefore, our major contribution is outlined as follows:

- Presented linear programming (LP) formulation for the selection of CHs.
- Presented LP formulation for data routing.
- GSA-based CH selection algorithm.
- Presented a cost function for the cluster formation.
- presented efficient encoding scheme for data routing.

The organization of the remaining part is presented in the following order. Section 2 presents the related works. An overview of GSA is provided in Sect. 3. The system models which are used in simulation are presented in Sect. 4. The Linear programming problem formulation is given in Sect. 5. The proposed GSA based algorithms are explained in Sect. 6. The simulation results are analysed in Sect. 7 followed by the conclusion in Sect. 8.

2 Related Works

In recent years, a few clustering and routing algorithms have been developed by using meta-heuristic approaches in WSNs. However, very few of them addressed hot spot problem. Therefore, we focus on nature inspired energy conserving techniques i.e., clustering and routing algorithms. A few nature inspired algorithms exist for clustering as well as routing in the survey. In [8,13], PSO based CH selection algorithms have been proposed. Thus, they do not mention how the clusters can form after CH selection is done. In [7], the authors have proposed two algorithms to address the hotspot problem. However, it may not solve hot

spot problem due to random selection of CHs and uncertainty of cluster formation. Jiang et al. [6] have developed PSO based energy balanced unequal clustering (EBUC) and also implemented routing algorithms. In [2], the authors have proposed an energy aware fuzzy approach for unequal clustering (EAUCF). However, in both [7] and [2], the SNs may be assigned to the nearest CHs that may lag to energy imbalance in the network. Moreover, complexity of determining routes is not considered. A fuzzy logic based unequal clustering algorithm has been developed (FBUC) [9] that does not consider the residual of CHs at the time of cluster formation that does not maintain the scalability of the network due to inefficient use of energy.

The proposed algorithms provide an energy efficient CH selection and energy efficient routing path in terms energy balancing by considering the minimum distance, residual energy and sink distance.

3 An Overview of GSA

The GSA [12] is a new meta-heuristic optimization algorithm based on the law of gravity, law of motion and mass interactions. In GSA, particles/agents/solutions in search space are considered as objects. The performance of these objects is measured as follows. The objects attract each other by the gravity force, and that causes a global movement of all objects towards the objects which have heavier masses.

The agents that hold heavy masses can be considered as good solutions. The agents with heavier mass could move more slowly than lighter ones. The lighter masses are known as worst solutions. Here, each agent has four specifications such as position, inertial mass, active gravitational mass and passive gravitational mass. The position of the mass corresponds to a solution of the problem, and its gravitational and inertial masses are determined using a predefined fitness function which determines the solution of a problem.

The N agents/particles (masses) defined in search space are as follows $P_i = [P_i^1, P_i^2, \dots, P_i^d, \dots, P_i^n]$ for $i = 1, 2, \dots, N$ where, P_i^d represents the i^{th} position agent in d^{th} dimension. The force acting on mass with N search space is defined as follows:

$$F_{ij}^d(t) = G(t) \frac{M_{pi}(t) \times M_{aj}(t)}{R_{ij}(t) + \epsilon} (P_j^d(t) - P_i^d(t)) \quad (1)$$

where M_{aj} is the active gravitational mass related to agent j , M_{pi} is the passive gravitational mass related to agent i , $G(t)$ is gravitational constant at time t , ϵ is a small constant, and $R_{ij}(t)$ is the Euclidian distance between two agents i and j . Equation 2 gives total force exerted by all agents on i^{th} at time t in d^{th} dimension:

$$F_i^d(t) = \sum_{j=1, j \neq i}^N rand_j F_{ij}^d(t) \quad (2)$$

where $rand_j$ is a random number in the interval $[0, 1]$.

In law of motion, the acceleration of an agent i at time t is defined as

$$a_i^d(t) = \frac{F_i^d(t)}{M_{ii}(t)} \quad (3)$$

where M_{ii} is the inertial mass of i^{th} agent.

Therefore, velocity and position of an agent i is updated using Eqs. 4 and 5 respectively

$$V_i^d(t+1) = rand_i \times V_i^d(t) + a_i^d(t) \quad (4)$$

$$P_i^d(t+1) = P_i^d(t) + V_i^d(t) \quad (5)$$

where $rand_i$ is a uniform random variable in the interval $[0, 1]$, it helps to randomized characteristic to the search.

The fitness/objective function is derived by defining the gravitational and inertia masses sing predefined defined fitness evaluation function. A heavier mass particle indicates better solution that has higher attractions and moves more slowly. Assuming the equality of the gravitational and inertia mass, the values of masses are calculated using the map of fitness. The update the gravitational and inertial masses as:

$$M_{ai} = M_{pi} = M_{ii} = M_i, i = 1, 2, \dots, N \quad (6)$$

and

$$m_i(t) = \frac{fit_i(t) - W(t)}{B(t) - W(t)}, M_i(t) = \frac{m_i(t)}{\sum_{j=1}^N m_j(t)} \quad (7)$$

where $fit_i(t)$ represent the fitness value of the agent i at time t , and, W is $worst(t)$ and B is $best(t)$ are defined as: $B(t) = \max_{j \in 1, \dots, N} fit_j(t)$ and $W(t) = \min_{j \in 1, \dots, N} fit_j(t)$.

4 System Models

Network model: In a WSN, we assume that the SNs become stationary after a random deployment of SNs. We consider that the SNs have the local information, like the distance of its neighbour SNs and also their residual energy levels as in [5, 10]. Similar to LEACH [5], the data gathering operation is divided into rounds. In each round, each SN collects the local data and sends it to its corresponding CH. Then, each CH aggregates the collected local data to discards the uncorrelated and redundant data and sends that aggregated data to the sink via few intermediate CHs. Every communication held through a wireless link.

Energy model: We use the same radio model for energy as considered in [5].

5 Problem Formulation

We first define some terminologies which are useful to present the *LP* formulation and proposed algorithms as follows.

1. $E(i)$ denotes the remaining energy of SN i
2. R_{max} is the maximum communication range of the SN
3. $dist(i, j)$ denotes the Euclidian distance between node i and node j
4. A set of SNs denoted by $S = \{S_1, S_2, \dots, S_n\}$
5. The set of *CHs* is defined as $\xi = \{C_1, C_2, \dots, C_m\}$ where $m < n$
6. $degree(C_i)$ is the number of SNs assigned to the CH C_i
7. $Com(C_i)$ is the set of *CHs*, that are in the communication range of C_i . The sink may also be a member of $Com(C_i)$ i.e.,
 $Com(C_i) = \{C_j \mid \forall C_j \in \xi + Sink \text{ and } dist(C_i, C_j)\}$
8. Ω_i is the set of SNs of that are within the communication range of CH C_i

5.1 LP Formulation for CH Selection

In case of CH selection, we consider the minimum distance, minimum sink distance and maximum residual energy of the nodes. The Linear Programming (LP) of the optimal *CH* selection, let it be L and is given as

$$\text{Maximize } L \quad (8)$$

Subject to

$$\text{Minimize } \sum_{j=1}^m \sum_{i=1}^{\Omega_j} dist(C_j, S_i) \quad (9)$$

subject to

$$dist(C_j, S_i) \leq d_{max} \mid S_i \in \Omega_j \quad (10)$$

$$dist(S_i, C_i) \leq d_{max}, \forall S_i \in S \text{ and } C_j \in \xi \quad (11)$$

$$E(C_j) > threshold \quad 1 \leq j \leq m \quad (12)$$

The constraint (10) states that the SN S_i is within the maximum communication range of C_j . The energy of C_j node must be greater than the threshold energy and it is stated in (11). In the constraint (12), χ_1, χ_2 and χ_3 are the control parameters of functions f_1, f_2 and f_3 respectively.

5.2 LP Formulation for Cluster Formation and Routing

Let d_i denotes the traffic load generated by the SN S_i and d_k^* be the routing load of cluster head C_k . Then the overall load W_j of CH C_j will be given as:

$$W_j = \sum_{i=1}^n d_i \times \alpha_{ij} + \sum_{i=1}^m d_m^* \times \beta_{ij} \quad (13)$$

where, α_{ij} and β_{ij} are the Boolean variables such that

$$\alpha_{ij} = \begin{cases} 1, & \text{if } S_i \text{ is assigned to } C_j \\ 0, & \text{Otherwise} \end{cases}$$

$$\beta_{kj} = \begin{cases} 1, & \text{if } C_k \text{ use } C_j \text{ as next hop CH} \\ 0, & \text{Otherwise} \end{cases}$$

Then the maximum load of the CHs is $W = \max\{W_i \mid \forall C_i \in \xi\}$. Now, we aim to minimize the overall maximum load of the CHs. Then, the linear programming of the clustering and routing problem can be derived as follows:

Minimize $W = \max W_i \mid \forall C_j \in \xi$

subject to

$$\sum_{j=1}^m \alpha_{ij} = 1 \mid \forall S_i \in S \quad (14)$$

$$\sum_{S_i \in S} d_i \times \alpha_{ij} + d_k^* \times \beta_{kj} \leq \mid \forall C_j \in \xi \quad (15)$$

$$\sum_{j=1}^m \text{dist}(i, j) \times \alpha_{ij} \leq d_{max} \mid \forall S_i \in S \quad (16)$$

$$\sum_{j=1}^m \text{dist}(C_i, C_j) \times \beta_{ij} \leq R_{max} \mid \forall C_i \in \xi \quad (17)$$

The constraint (14) expresses that an SN can be assigned to one and only one CH. The constraint (15) ensures that the load of a CH must not exceed the overall maximum load of the CHs. The constraint (16) indicates that the SNs are assigned to the CH and the constraint (17) ensures that the CH C_k must be within the communication range of the CH C_j and C_k is producing routing load on C_j .

6 Proposed Method

The basic purpose of the proposed clustering algorithm is to choose K number of CHs in n SNs. The selection of the CHs depend on the residual energy of the SNs, distance from their possible CHs and from the sink. Once the CHs are chosen, the SNs can be assigned to their closest CHs to form the cluster.

6.1 A GSA Based Energy Efficient Cluster Head Selection Algorithm

Here, we describe our proposed approach to select K number of CHs from n number of SNs using GSA. Let GSA has R agents that are considered to

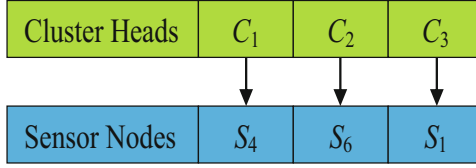


Fig. 1. Initialization of an agent

construct R solutions. Each agent forms a solution string of length K , in which each element represents a CH .

Initialization of an agent: Let $A_i = [x_i^1(t), x_i^2(t), x_i^3(t), \dots, x_i^D(t)]$ be the i^{th} agent i.e., solution. Each component $x_i^d(t)$ maps an SN to be selected as a CH where $1 \leq i \leq N_A$, $1 \leq d \leq D$. Each component is initialized by a randomly generated number between 1 and 2. The component of d^{th} dimension of an agent maps to the SN S_d as a CH . As an example, for a set of 10 SN and 3 CHs , a solution string is shown in Fig. 1, in which S_4, S_6 and S_1 are selected as C_1, C_2 and C_3 respectively.

Derivative function for cluster head selection: We derive the fitness function based on three objective functions. The reciprocal of minimum total distances between each SN has been taken for maximizing the objective function. In other words,

$$f_1 = \frac{1}{dist(S_j, S_l)} \forall S_j \in Com(S_l) \text{ and } S_l \in \xi \quad (18)$$

The maximum remaining energy of the corresponding SN amongst the neighbour SNs has been considered to maximize the objective function. In other words

$$f_2 = \frac{E(S_i)}{\sum E(S_i)} \forall S_i \in Com(S_j) \text{ and } S_i \in S \quad (19)$$

The objective function, which must be maximized, is the reciprocal of minimum sink distance of corresponding sensor node. In other words,

$$f_3 = \frac{1}{dist(S_i, Sink)} \quad (20)$$

Here, the weight value χ_i can be multiplied by each objective function and converted into a single objective function for clustering. In other words, $F_1 = \chi_1 \times f_1 + \chi_2 \times f_2 + \chi_3 \times f_3$ where χ_1, χ_2 and χ_3 are the control parameters (weights) of the functions f_1, f_2 and f_3 respectively such that $\chi_1 + \chi_2 + \chi_3 = 1$. The detailed algorithm is presented in Algorithm 1.

Algorithm 1. GSA based CH selection algorithm

Input: 1. $S = S_1, S_2, S_3, \dots, S_n$
 2. Swarm agents of size N_A
 3. Number of dimensions of an agent = Number of CHs = m
Output: $\xi = C_1, C_2, C_3, \dots, C_m$
 Step:1 Initialize agents $A_i, \forall i, 1 \leq i \leq N_A, A_i = [x_i^1(t), x_i^2(t), x_i^3(t), \dots, x_i^D(t)]$,
 D is number of CHs, $C_k = Index(Com(S_d)) \times x_i^d \times m$
 Step:2 **for** ($t=1$ to Terminate) **do**
 for ($i=1$ to N_A) **do**
 2.1 Compute fitness (A_j)
 2.2 Update best and worst fitness of all agents
 2.3 Calculate $M_i(t)$ and $x_i^d(t)$ of each agent
 2.4 Update **velocity** and **position** of A_i
 Step:3 Stop

6.2 A GSA Based Energy Efficient Cluster Formation and Routing Algorithms

In cluster formation, the normal SNs are assigned to their nearest CHs as shown in Fig. 2. Then, the GSA routing algorithm works in two phases: Neighbour CH discovery and routing which are described subsequently as follows.

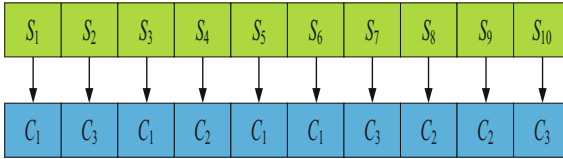


Fig. 2. Illustration of cluster formation

Initialization of routing agent: Let $B_i = [y_i^1(t), y_i^2(t), y_i^3(t), \dots, y_i^D(t)]$ be the i^{th} routing agent. Each component $y_i^D(t)$ maps a CH to be selected as a next hop CH and $1 \leq i \leq N_B, 1 \leq d \leq D$. The component of d^{th} dimension of an agent, i.e., $y_i^D(t)$ maps to the CH C_d as a next hop CH. As an example, a routing solution is shown in Fig. 3.

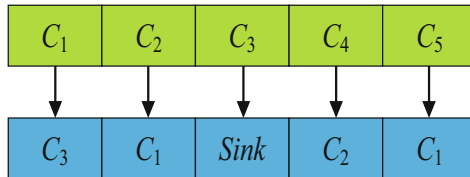


Fig. 3. Initialization of a routing agent

Neighbour CHs discovery: During this process, broad casted packets are exchanged between the *CHs* which contain the information of the *CH* such as residual energy of the *CH* and its distance from the sink. These messages mainly provide an indexing for routing. The source *CH* C_i discovers the neighbour *CHs* which are within its communication range. Each intermediate *CH* forwards the request only to the neighbours. Thus, the *CH* C_i request is sent only to a neighbour C_j which satisfies the condition, $dist(C_i, sink) \geq dist(C_j, sink)$. In routing phase, when C_j sends data towards *sink*, C_j chooses the next hop *CH* C_i for the data transmission according to the following derived function. Note that if the sink is within the communication range of *CHs*, those *CHs* choose the sink as next hop and transmit the data to the sink directly.

6.2.1 Derivative Function for Routing

In the proposed routing algorithm, we have three objectives that are discussed as follows. A *CH* C_l should select that *CH* C_k from its neighbour (next hop) *CHs* which has maximum residual energy. In other words,

$$f_4 = \frac{E(C_k)}{\sum E(C_l)} \forall C_i \in Com(C_k) \quad (21)$$

During transmission of data, a significant quantity of energy is consumed. Energy consumption by the *CH* is increased with the increase of transmission distance. So, a *CH* C_l selects the nearest *CH* C_k from its neighbour *CHs* for forwarding data. The shorter the distance between C_l and C_k , the higher is the chance of selecting C_k as next hop *CH*. As we need to maximize objective function, we consider reciprocal of the minimum Euclidean distances between each *CH*.

$$f_5 = \frac{1}{dist(C_k, C_l)} \forall C_l \in Com(C_k) \text{ and } C_k \in \xi \quad (22)$$

If a next hop *CH* of C_l is far away from the sink, then the data packets have to travel longer path to reach the sink thereby consuming more energy for longer distance communication with the sink. So, a *CH* C_l should select that *CH* C_k which is near to the sink. Therefore, the objective function, which must be maximized, is the reciprocal of minimum sink distance of corresponding *CH*.

$$f_6 = \frac{1}{dist(C_k, Sink)} \quad (23)$$

We now combine all the above objective functions, f_4 , f_5 and f_6 of Eqs. 21, 22 and 23 respectively to convert them into an objective function to yield the routing function as follows:

$$F_2 = \chi_4 \times f_4 + \chi_5 \times f_5 + \chi_6 \times f_6 \quad (24)$$

where, χ_4 , χ_5 and χ_6 are the control parameters (weights) of the functions f_4 , f_5 and f_6 respectively such that $\chi_4 + \chi_5 + \chi_6 = 1$.

Algorithm 2. GSA based routing algorithm

Input: 1. Set of CHs $\xi = C_1, C_2, C_3, \dots, C_m, sink$

2. Swarm agents of size N_B

3. Number of dimensions of an agent = m

Output: Routing paths for all CHs

Step:1 Initialize agents $B_i, \forall i, 1 \leq i \leq N_B, B_i = [y_i^1(t), y_i^2(t), y_i^3(t), \dots, y_i^D(t)]$,
 D is number of CHs, $C_k = Index(Com(C_d)) \times y_i^d \times m$

Step:2 $Com(C_i) = \emptyset$

Step:3 **while** (C_i receiving HELLO messages from C_j) **do**

 3.1 $Com(C_i) = Com(C_i) \cup C_j$

 3.2 $NextHop(C_i) = \{C_j \mid \forall C_j \in Com(C_i) \text{ and } dist(C_j, Sink) < dist(C_i, Sink)\}$

Step:4 **while** ($NextHop(C_i) \neq NULL$) where $i = 1$ to m **do**

 4.1 Compute fitness (B_i)

 4.2 Update *best* and *worst* fitness of all agents

 4.3 Calculate $M_i(t)$ and of each agent

 4.4 Update **velocity** and **position** of B_i

Step:5 Stop

7 Simulation Results

We substantially experimented the proposed algorithms through simulation in C programming language and MATLAB (version 7.5) on an Intel Core i7-2600 processor and 2 GB RAM running on the operating system Microsoft Windows 7 professional. The simulations were presented with two different scenarios i.e., by placing the sink at (250, 250) and (0, 250) as *WSN#1* and *WSN#2* respectively in the target area of size $500 \times 500 \text{ m}^2$. The parameters were set as: number of *SN* 200–800, initial energy 0.5 J, packet length 4000 bits, message length 500 bits, $d_0 = 60 \text{ m}$, $efs = 10 \text{ pJ/bit/m}^2$ and $emp = 0.0013 \text{ pJ/bit/m}^4$.

Network lifetime: Here, we present the experimental results of the proposed algorithms in terms of network lifetime which we consider that lifetime is the number of rounds until the first node dies. The results are shown in Fig. 4. It is obvious to note that the proposed algorithms outperform than nCRO, FBUC and PSO. The reason is that the proposed algorithm considers three objectives in CH selection and routing phases i.e., minimum total distance between SNs/CHs, energy of SNs/CHs and the sink distance from the SN/CH.

Number of alive sensor nodes: Now, we perform the comparison of the proposed algorithms with the existing algorithms in terms of number of alive SNs against number of rounds for both the scenarios as shown in Fig. 5. The simulation results show that the proposed algorithms outperform compared algorithms.

Residual energy: We plot the results of number of live sensor nodes against the simulation rounds for both the scenarios *WSN#1* and *WSN#2* for 200 SNs and compare the results of existing algorithms which is shown in Fig. 6. Here, we

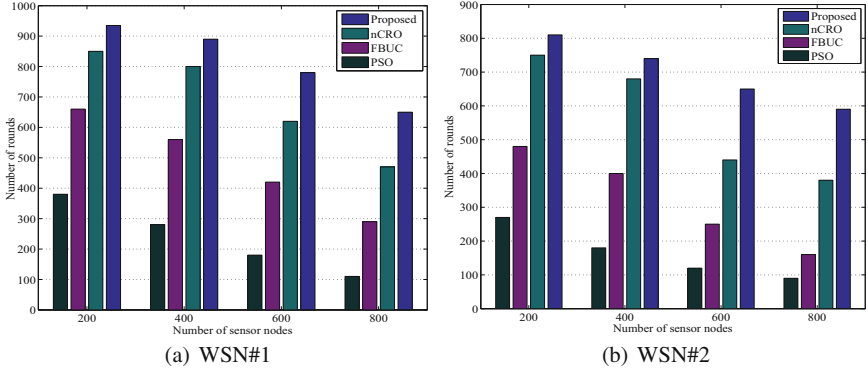


Fig. 4. Comparison of network lifetime

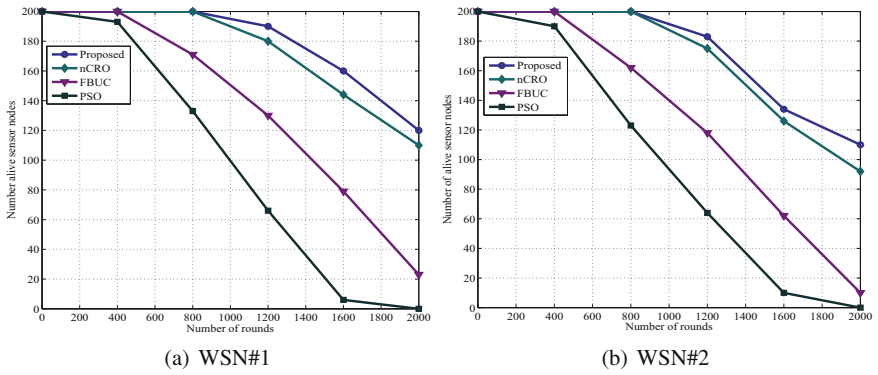


Fig. 5. Comparison in terms of alive sensor nodes

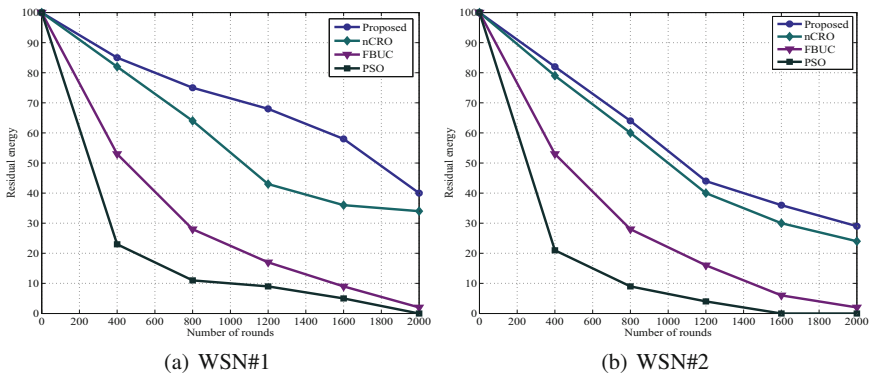


Fig. 6. Comparison in terms of residual energy

observe that proposed algorithms perform better than the existing algorithms and also WSN#1 results are better than the scenario WSN#2.

8 Conclusion

In this paper, we presented LP formulations for CH selection, cluster formation and routing. Next, we presented GSA based energy efficient clustering and routing algorithms by deriving novel fitness functions with efficient encoding schemes. In derivation of fitness functions of CH selection and routing, we have considered various parameters such as Euclidian distance from the SNs to CHs, CHs to the sink and energy of SNs and CHs. The proposed algorithms are simulated over two different scenarios of WSNs to show the superiority of the proposed algorithms. Those results are also compared and demonstrated their efficacy over various existing algorithms such as nCRO, FBUC and PSO in terms of network lifetime, energy consumption and alive SNs.

References

1. Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless sensor networks: a survey. *Comput. Netw.* **38**(4), 393–422 (2002)
2. Bagci, H., Yazici, A.: An energy aware fuzzy approach to unequal clustering in wireless sensor networks. *Appl. Soft Comput.* **13**(4), 1741–1749 (2013)
3. Banka, H., Jana, P.K., et al.: PSO-based multiple-sink placement algorithm for protracting the lifetime of wireless sensor networks. In: *Proceedings of the Second International Conference on Computer and Communication Technologies*, pp. 605–616. Springer (2016)
4. Guo, W., Li, J., Chen, G., Niu, Y., Chen, C.: A PSO-optimized real-time fault-tolerant task allocation algorithm in wireless sensor networks. *IEEE Trans. Parallel Distrib. Syst.* **26**(12), 3236–3249 (2015)
5. Heinzelman, W.B.: Application-specific protocol architectures for wireless networks. Ph.D. thesis, Massachusetts Institute of Technology (2000)
6. Jiang, C.J., Shi, W.R., Tang, X.L., et al.: Energy-balanced unequal clustering protocol for wireless sensor networks. *J. China Univ. Posts Telecommun.* **17**(4), 94–99 (2010)
7. Kuila, P., Jana, P.K.: Energy efficient clustering and routing algorithms for wireless sensor networks: particle swarm optimization approach. *Eng. Appl. Artif. Intell.* **33**, 127–140 (2014)
8. Latiff, N.A., Tsimenidis, C.C., Sharif, B.S.: Energy-aware clustering for wireless sensor networks using particle swarm optimization. In: *IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC 2007*, pp. 1–5. IEEE (2007)
9. Logambigai, R., Kannan, A.: Fuzzy logic based unequal clustering for wireless sensor networks. *Wirel. Netw.* **22**(3), 945–957 (2016)
10. Ok, C.S., Lee, S., Mitra, P., Kumara, S.: Distributed energy balanced routing for wireless sensor networks. *Comput. Ind. Eng.* **57**(1), 125–135 (2009)
11. Rao, P.S., Banka, H.: Novel chemical reaction optimization based unequal clustering and routing algorithms for wireless sensor networks. *Wirel. Netw.* 1–20 (2016)

12. Rashedi, E., Nezamabadi-Pour, H., Saryazdi, S.: GSA: A Gravitational Search Algorithm. *Inf. Sci.* **179**(13), 2232–2248 (2009)
13. Singh, B., Lobiya, D.K.: A novel energy-aware cluster head selection based on particle swarm optimization for wireless sensor networks. *Hum. Centric Comput. Inf. Sci.* **2**(1), 13 (2012)

Web Service Recommendation Based on Semantic Analysis of Web Service Specification and Enhanced Collaborative Filtering

S. Subbulakshmi¹(✉), K. Ramar², Ameena Shaji¹,
and Parvathy Prakash¹

¹ Department of Computer Science and Applications,
Amrita School of Engineering, Amritapuri, Amrita Vishwa Vidyapeetham,
Amrita University, Amritapuri, Kollam, India
subbulakshmis@am.amrita.edu,
ameenashaji.009@gmail.com,
parvathiprakashlll@gmail.com

² Department of CSE, Einstein College of Engineering, Tirunelveli, India
kramar.einstein@gmail.com

Abstract. With growing momentousness of Internet applications, digital world is overwhelmed with huge number of web services. To ease the job of selecting relevant WS in service composition process, recommendation system of Web Services is designed. It uses semantic analysis of WS along with enhanced collaborative filtering. Ontology based Semantic Analysis performed using Tversky Content Similarity Measure helps to identify most similar functionally relevant WS. The collaborative filtering process uses DBSCAN clustering and PCC similarity to identify highly collaborative WS, based on ratings given by experienced users. To overcome the existence of sparse data in WS ratings and to enhance filtering process, SVM Regression is implemented before collaborative filtering. Relative frequency method is applied to amalgamate collaborative and semantic similarity values of WS. The methodology is proved to produce more realistic, accurate and efficient WS recommendation. Future focus may be towards knowledge based filtering with real world contextual information.

Keywords: WS ontology description · Semantic content filtering · Enhanced collaboration · Efficient filtering of WS

1 Introduction

With the beginning of 21st century, the Internet began developing and expanding with magnificent velocity. The amount of information available on the Internet has got extremely prodigious, resulting in an information overflow. Performing complex business operations and providing quality information to the end users can be achieved through the use of selection of most relevant Web Services. Existing systems [10] are not enough efficient to retrieve the actual web services desired by the user up to a limit.

Contextual based Recommender systems [9] are competent of solving this issue up to a great extent by selecting the services in par with the requirements of the end users. It has emerged as a powerful tool for reducing the complexity of information/services enormity. Web service recommendation is the process of identifying the measure of usefulness of the web services and proposing them to the user.

Web services provide a regulated manner to incorporate web applications using open standards over an Internet protocol back bone, using some platform and language independent interfaces meant for easily assimilating heterogeneous systems. Web Services provide interoperability between various applications. UDDI, WSDL and SOAP define standards for service discovery, description, and messaging protocols for web services respectively.

Many researchers are focused in adopting the content based and collaborative filtering approach in the process of selecting the web services. Collaborative filtering [15] mainly centres on identifying neighbourhoods of target user consisting of other users with similar interests or preferences. Collaborative recommenders rely on user profiles, usually represented as rating vectors. Examples of such applications includes recommending movies, tour destinations, music, games etc.

Existing content based filtering methods employs the exact keyword similarity measures for the selection of web services. Majority of the traditional methodologies focus on searching the existing UDDI registries or implements keyword based search process. This resulted in poor recommendation and also requires clear and correct queries from the user. Therefore, in this paper, we present a high performance recommender system of web services based on user preferences, which makes use of machine learning and data mining techniques like regression [2] and clustering coupled with the advantages of semantic analysis. It is able to provide the end user with the most relevant web service which delivers the most pertinent information.

An enhanced content and collaborative filtering approaches for web services is designed to select the most appropriate services which handles data sparsity, data overload and scalability issues of the existing system. The initial phase of the approach concentrates on content based filtering along with semantic analysis. It consists of two central tasks such as domain feature's similarity checking, and matching of input output parameters. An ontology is designed using necessary data extracted from the corresponding WSDL files of web services. Semantic based similarity calculations are performed on domain features and input-output parameters of the given web services using Tversky's Content Similarity Measure to reach at a set of highest associated web services.

Secondary phase of the approach mainly encloses three major stages - Data sparsity removal, clustering of similar items, and ranking of similar web services. Unrated or unobserved web services available in web may cause data sparsity. In order to tackle with the problems of data sparsity, we use SVM regression to fill in missing user ratings. Grouping of similar web services is achieved through DBSCAN. When a user inputs a search query, its corresponding cluster is identified and web services that fall on the identified cluster are then ranked using PCC.

Tertiary phase focuses on combining outputs from previous filtering modules to constitute an improved and more accurate high quality recommender output. Relative frequency method is implemented for this purpose. The efficient enhanced

recommendation of web services can be widely used in the service composition where the service broker agent wants to automate the dynamic selection of the best services from the existing set of web service registries.

The rest of paper includes: the detailed study of the existing research work in Sect. 2 Related Works, illustration of the proposed Enhanced Recommendation System of WS in Sect. 3 System Methodology, the analysis report of Enhanced Collaborative Filtering and Sematic based Content Filtering approach in Sect. 4 Result and Analysis and the concluding part with directions for future improvements in Sect. 5 Conclusion.

2 Literature Review

A number of researches have already been done on recommendation system for web services. Yang et al. [3] gives the semantic similarity between web services through calculating the normalized google distance. It uses google massive terms and open google search engine to determine the normalized google distance between notions.

Lina Yao proposes an approach [1] that joins both Content based and Collaborative based methodologies by considering both appraisals and functionalities of web administrations utilizing a Probabilistic Generative Model. The idle inclinations are measurably assessed utilizing Expectation-boost calculation. To overcome information sparsity issue, data smoothing method is adapted. The system is further improved by content similarity and implicit user description aspect model.

An implementation of automated adaptive framework [5] for the WS coupled with optimisation of QoS based on quality specifications in the Web Services Ontology. Using this framework, the users are able to acquire a set of web services, by consuming the context information of users and services, which is further enhanced by the Quality factors of those web services.

Mingxin Gan proposes an approach [7] that relies on ontologies to determine the semantic similarity between tags. The system uses five categories of methods based on semantic distance, information content, properties of tags, ontology hierarchy [15], and hybrid methods. Semantic similarity is calculated by the length of the path from the leaf nodes to the root node. Tags are represented as collection of features, normalization and set theory functions are applied to estimate semantic similarity between tags.

An interactive composition approach [4] by Evren Sirin, using matchmaking algorithms is presented to help users to filter and select services while building the composition. The filtering and selection of services helps the users in the composition process. A travel recommendation system in Semantic Web using Ontology is designed [6] by Chang Choi. The Metadata is made by preference profile and transaction profile. The Travel Ontology is made by OWL Rule based on Description Logic.

3 Proposed Methodology

Every service available on Internet are provided by different web services. The core area of our work is an improved web service recommendation system that recommends the most relevant information available on web. The proposed methodology shown in Fig. 1

relies on two major phases Sematic Based Content (SCB) filtering and Enhanced Collaborative filtering (ECR) methods for the selection of relevant service. The key idea of our proposed approach is to recommend web services by selecting the semantically similar service with high user ratings provided by different users.

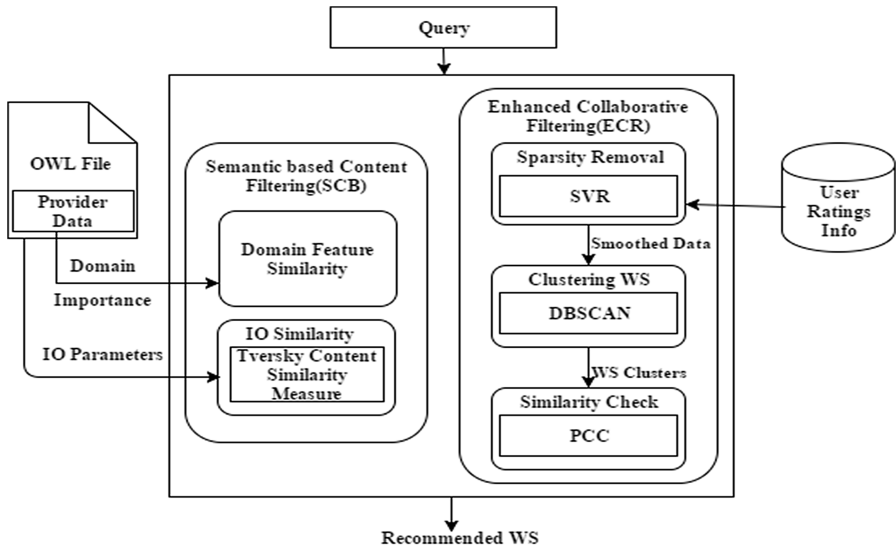


Fig. 1. Methodology for Web Service Recommendation

The design methodology implements three major features: removal of sparse data using regression methodology, improve efficiency by clustering of data and to accomplish more realistic selection by adopting the semantic based methodology. An Ontology of the web services is maintained as a repository to store the details of the services and their relationships to be used for the sematic retrieval of the required services. In essence, recommendation is based on an automatic dynamic selection of pertinent services, subject to the filtering process of web services collaboration based on the ameliorated ranking given by multiple users which exhibit similar preferences or behaviours.

3.1 Semantic Based Content Filtering (SCB)

According to W3C, the semantic web establishes a standard framework which enables to share or reuse data and services across application, organizations, enterprise, and community boundaries. Initial phase of our work is centred on Semantic Content based filtering approach which uses the ontology for identifying the web services with the required specifications given in the user query. Each web service consists of a WSDL file which defines how a service can be called, parameters required as its input/output data, domain name and other specifications. By consuming the materials available in the WSDL file, an ontology is populated with OWL, the ontology language for semantic web.

SCB similarity is calculated by using improved Tversky’s Content Similarity Measure considering domain features and input-output parameters of services. Firstly,

the domain type and prominence of the service are expanded to find the similar services. It is followed by an input-output parameter matching to analyse the similarity between user requirements (as query) and WS message descriptions. Thus, web services with highest similarity values are selected to produce the output of SCB filtering.

3.1.1 Web Service Ontology Creation

The SCB filtering starts with the creation of an ontology for web services. Ontology is the working model of objects belonging to a particular area of interest and their semantic relationships to each other. Ontologies are authored using Ontology Web Language (OWL), a set of knowledge representation languages built upon a W3C XML standard. Our system relies upon a web service ontology developed using Protégé and it defines a set of web services and their mutual semantic relationships. The specification of web services given by service providers residing in service registries are retrieved, to be used in the process of web services ontology creation.

Ontology created includes a set of classes and subclasses which depicts the relationship between different services. The service description like domain name, input/output parameters and importance of the services are included as data properties in the ontology. Figure 2 shows the sample of the ontology created in the system using protégé tool. For the selection of required services, the semantic details of web services are retrieved to calculate the similar web services.

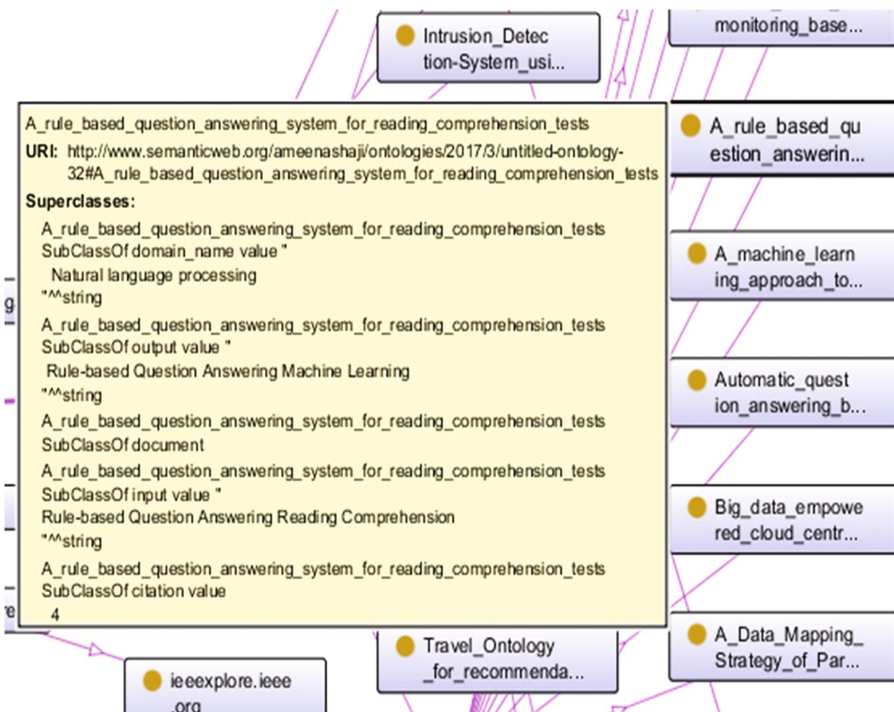


Fig. 2. Ontology of web services with service description

Similarity between user requirements and web service description is calculated using improved [7] Tversky Content Similarity Measure. Tversky defines a similarity measure according to the matching process, which generates a similarity value based on, not only common factors but also distinct features of web services.

3.1.2 Content Similarity Measure

This approach is an efficient method used in determining information-theoretic similarity values. Unlike other models this model determines the similarity between the matching-features and then it evaluates the impact of non-matching features of those web services to assess the similarity between them.

Algorithm to calculate the similarity between given web service W_k with all other web services in the ontology:

Improved Tversky’s Similarity for WS

1. Identify the domain name d_k and set of properties p_k of the web service W_k
 2. Select the set of web service S whose domain name is equal to d_k
 3. For every web service W_i in S , do the following:
 - a Retrieve the input/output property p_i of web service W_i
 - b Calculate the Improved Tversky’s similarity $IT_{Sim(i)}$ between p_k and p_i ,

$$T_{Sim(w_i,w_k)} = IT_{Sim} + C_{Sim}$$
 - c Retrieve and add the importance value I_i of W_i with $T_{Sim(i)}$.
-

3.1.3 Improved Tversky Content Similarity Measure

The improved similarity method implements the Tversky’s normalization with the set-theory functions intersection ($p_i \cap p_k$), difference (p_i / p_k) and the Cosine similarity functions. The standard formulation is given as:

$$IT_{Sim} = \frac{|p_i + p_k|}{|p_i \cap p_k| + \mu|p_i/p_k| + (\mu - 1)|p_k/p_i|} \tag{1}$$

$$or\ 0 \leq \mu \leq 1$$

$$C_{Sim} = \frac{p_i \cap p_k}{p_i + p_k - (p_i \cap p_k)} \tag{2}$$

where p_k and p_i corresponds to the description sets of web service W_k and W_i and μ is a function [7] that defines the relative importance of the non-common features. The semantic relationships maintained in ontology is used to determine the relative weightage for properties of web services. Thus, by dynamically assigning accurate value for μ and by aggregating the importance of the service, this method is able to select content based similar web services.

3.2 Enhanced Collaborative Filtering (ECR)

Generally Collaborative filtering is used in recommendation system to identify popular items among peer users with the help of the ratings given by different users. We have adopted an Enhanced Collaborative (ECR) filtering process, which employs sparsity removal and clustering methods to select popular web services. Sparsity removal is used to fill the missing values in the user ratings data set with SVM regression methods. DBSCAN clustering method is implemented to group related services so that only those services with basic features could be considered for collaborative calculation.

3.2.1 SVM Regression

SVM Regression is used for predicting missing values. Rated web services are contained in trained file and unrated web services are contained in tested file. Unrated web services in the test file are rated using ratings contained in train file. SVR derives a function $f(x)$ which has less deviation between observed and predicted training samples. It also minimizes the error which is a combination of training error and a regularization term that controls the complexity of the hypothesis space.

Algorithm for Sparsity Removal

1. Actual dataset with user ratings is divided into training set and testing set.
 2. Data set is classified and predicting using SMOreg classification and prediction.
 3. Classification is done using *classifier.getClassification* and class labels are assigned.
 4. Maximum marginal plane is identified in order to maximize the prediction accuracy.
 5. Real values are predicted to fill the sparse data with *evaluation.NumericPrediction*
-

3.2.2 DBSCAN

Density-Based Spatial Clustering groups all services of data set into service of clusters and noise. The key idea of clustering is to identify whether the minimum number of services are present within the given radius i.e., the density in the neighbourhood has to exceed some threshold value.

Input: User ratings filled using SVM regression and Eps value.

Algorithm for Clustering Web Services

1. Select an arbitrary web service w_a
 2. Recognize all web service's density reachable from web service w_a with Eps and MinPts
 3. If w_a is a core service, a cluster is formed.
 4. If w_a is a border service, no web services are density reachable from w_a and DBSCAN visits the next web service of the dataset.
 5. Continue the process until all of the web services have been processed.
-

DBSCAN applies Euclidean distance to find distance between two web services. If the Cartesian coordinates of user ratings for two web services are $w_a = (w_{a1}, w_{a2}, \dots, w_{an})$ and $w_b = (w_{b1}, w_{b4}, \dots, w_{bn})$ in Euclidian n space, the distance (d) from w_a to w_b , or from w_b to w_a is given by the Pythagorean rule:

$$d(w_a, w_b) = \sqrt{\sum_{i=1}^n (w_{ai} - w_{bi})^2}. \tag{3}$$

where n is the number of users and w_{ai} is the rating given by i^{th} user to web service w_a and also w_{bi} is the rating given by i^{th} user to web service w_b .

3.2.3 Pearson Correlation Coefficient

The final step Collaborative Filtering process is PCC which is a quite famous algorithm used for selection of candidate items. PCC measures the strength of linear association between two variables, where $r = 1$ means a perfect positive correlation and the value $r = -1$ means a perfect negative correlation. The selection of the highly collaborative web services among the cluster is effectively computed using this algorithm.

The Correlation between the web service W_k queried by the user with all other web services W_i in the cluster, to which the queried web service belongs is implemented using PCC equation:

$$Cor(W_k, W_i) = \frac{\sum_{j=1}^n (W_{kj} - \bar{w}_k)(W_{ij} - \bar{w}_i)}{\sqrt{\sum_{j=1}^n (W_{kj} - \bar{w}_k)^2} \sqrt{\sum_{j=1}^n (W_{ij} - \bar{w}_i)^2}}. \tag{4}$$

where W_{kj} and W_{ij} are the ranks given by n users for web service W_k and W_i .

4 Result and Analysis

The simulation of the system is tested with the sample data set which includes the ontology owl file with web service descriptions and the ratings csv file with the user ratings provided by different users for a set web services. The online user should rate each web service available according to the level up to which he/she is satisfied with that WS. Ratings can range from one to five. Unobserved user ratings are assumed as 0, i.e. the sparse data. Only those sparse data are filled up using SVM Regression, it is done as the pre-processing step. Sample results of filling the sparse data is shown in Tables 1 and 2.

Table 1. Sample user ratings data set with sparse data

WS/ User ID	Admin@ .com	aishu@ .com	ajith@ .com	akhil@ .com	akhilv@ .com	akshaya@ .com	amee@ .com
w1	4	4	3	3	4	0	0
w2	3	3	4	5	3	0	0
w3	0	2	1	4	0	0	0
w4	2	1	2	3	2	1	2
w5	0	0	4	1	0	3	2

Table 2. Sample user rating data set with filled values after pre-processing

WS/ User ID	Admin@ .com	aishu@ .com	ajith@ .com	akhil@ .com	akhilv@ .com	akshaya@ .com	amee@ .com
w1	4	4	3	3	4	5	1
w2	3	3	4	5	3	4	1
w3	4	2	1	4	4	1	1
w4	2	1	2	3	2	1	2
w5	1	3	4	1	3	3	2

Table 3. Results of SCB and ECR filtering.

WS	TSim	CSim	ITSim	ECR	Nor. ECR	RF
w1	1.376	0.153	3.764	0.22	2.361	0.139
w2	1.4	0.166	1.783	0.209	2.284	0.069
w3	1.506	0.222	1.864	0.598	5	0.086
w4	1.764	0	2.882	0.025	1	0.073
w5	1.636	0.265	4.961	0.123	1.684	0.178
w6	1.733	0.3	4.016	0.355	3.304	0.186
w7	1.969	0.3	2.134	0.377	3.457	0.087
w8	2.173	0.25	4.211	0.095	1.489	0.15

The results of semantic based and collaborative filtering are shown in Table 3. Relative Frequency (RF) method is used for combining the SCB and ECR filtering outputs together for a better recommender result. The proportion of total possible number of events to the count of the favourable events is termed as relative frequency. Relative Frequency (RF) method is used for combining the SCB and ECR filtering outputs together for a better recommender result. The proportion of total possible number of events to the count of the favourable events is termed as relative frequency. RF functions for Similarity and Collaborative filtering are given by:

$$Relative\ Frequency_{(w_k)} = \frac{(Sim_i + Cor_i)}{\sum_{i=1}^n (Sim_i + Cor_i)}. \quad (5)$$

where Sim_i is the similarity score, Cor_i is the correlative score of the k^{th} web service and n is the number of web services. The above function denotes relative frequency as a proportion.

The above table shows the detailed result of SCB and ECR filtering and the integrated results using Relative Frequency Method. Moreover, graphical representation of the above results are also shown in Figs. 3 and 4.

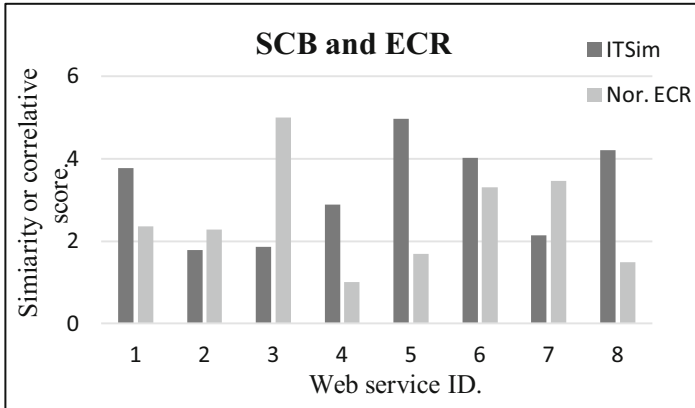


Fig. 3. Similarity values of SCB and ECR of web service.

The above graph reveals that the services are given higher ranks only when both SCB and ECR values are relatively higher as Web Service 6, 5, 8. If the services fails to score higher values in any of the filtering methods, they are given least importance in the process of selection like services 2, 4 shown in the above graph.

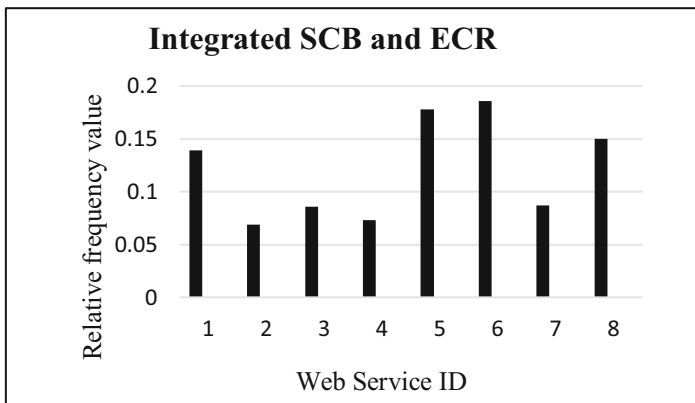


Fig. 4. Integrated Outcome of SCB and ECR filtering by applying RF.

The final ranking and recommendation of web services with regard to the given web service is shown in Fig. 5. The results of the implemented system revealed to be closest to the user expectation as the results produced is better than recommendation made without sparsity removal. The execution time of the system is reduced to greater extent as the clustering methodology is performed before PCC for web services.



Fig. 5. Ranking of Web Services.

The elaborated web service selection process with the semantic and enhanced collaborative filtering methodology could be used in the process of service collaboration by broker services for selecting the most appropriate services in par with their requirements.

5 Conclusion

The process of selecting and recommending relevant web services from a wide variety of available choice is an area of concern in Service Oriented Computing. Most current recommendation approaches focus on either UDDI registries, or keyword-dominant, QoS-based Web service search engines that have limitations such as reduced recommendation performance and dependence on detailed, precise search queries from the user. Our combined approach simultaneously considers user ratings similarities along with semantic content of Web services. As the methodology also considers filling of missing user ratings, the recommender output is much better with improved accuracy. Our approach only considers web services semantics. Semantic analysis of keywords can be incorporated as a part of future advancement.

References

1. Yao, L., Sheng, Q.Z., Ngu, A.H.H., Yu, J., Segev, A.: Unified collaborative and content-based web service recommendation. *IEEE Trans. Serv. Comput.* **X(X)** (2014)
2. Song, X., Zhou, T., Zhang, H.: Support vector regression estimation based on non-uniform lost function. In: *IEEE Engineering in Medicine and Biology 27th Annual Conference, Shanghai, 2005*, pp. 1127–1130 (2005)

3. Yang, H., Fu, P., Yin, B., Ma, M., Tang, Y.: A semantic similarity measure between web services based on google distance. In: 35th IEEE Annual Computer Software and Applications Conference, pp. 14–19 (2011)
4. Sirin, E., Hendler, J., Parsia, B.: Filtering and selecting semantic web services with interactive composition techniques. *IEEE Intell. Syst.* **19**, 42–49 (2004). doi:[10.1109/MIS.2004.27](https://doi.org/10.1109/MIS.2004.27)
5. Subbulakshmi, S., Ramar, K., Renjitha, R., Sreedevi, T.U.: Implementation of adaptive framework and WS ontology for improving QoS in recommendation of WS. In: Rodriguez, J.C., Mitra, S., Thampi, S., El-Alfy, E.S. (eds.) *Intelligent Systems Technologies and Applications 2016. ISTA 2016. AISC*, vol. 530, pp. 383–396. Springer, Cham (2016). doi:[10.1007/978-3-319-47952-1_30](https://doi.org/10.1007/978-3-319-47952-1_30)
6. Choi, C., Cho, M., Kang, E., Kim, P.: Travel ontology for recommendation system based on semantic web. In: Conference: Advanced Communication Technology, ICACT March 2006, pp. 624–627 (2006)
7. Gan, M., Dou, X., Jiang, R.: From ontology to semantic similarity: calculation of ontology-based semantic similarity. *Sci World J* 2013, Article ID 793091 (2013)
8. Hu, Y., Peng, Q., Hu, X., Yang, R.: Time aware and data sparsity tolerant web service recommendation based on improved collaborative filtering. *IEEE Trans. Serv. Comput.* **8**(5), 782–794 (2015)
9. Basu, C., Hirsh, H., Cohen, W.: Recommendation as classification: using social and content based information in recommendation. In: *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, pp. 714–720 (1998)
10. Smita, A.: Recommender system review. *Int. J. Comput. Appl.* **71**(24), 0975–8887 (2013)
11. Bahramy, F., Crone, S.F.: Forecasting foreign exchange rates using support vector regression. An empirical evaluation of mean reversion using bollinger bands. In: *IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFER)*, pp. 34–41 (2013)
12. Yao, L., Sheng, Q.Z., Segev, A., Yu, J.: Recommending web services via combining collaborative filtering with content-based features. In: *IEEE 20th International Conference on Web Services*, pp. 42–49 (2013)
13. Maximilien, E.M., Singh, M.P.: A framework and ontology for dynamic web services selection. *IEEE Internet Comput.* **8**(5), 84–93 (2004)
14. Lecue, F.: Combining collaborative filtering and semantic content based approaches to recommend web services. In: *Proceedings of the 4th IEEE International Conference on Semantic Computing (ICSC 2010)*, Pittsburgh, PA, USA (2010)
15. Linden, G., Smith, B., York, J.: Amazon.com recommendations item-to-item collaborative filtering, pp. 76–80. *IEEE Computer Society*, Washington, DC (2003)
16. Sajeev, G.P., Ramya, P.T.: Effective web personalization system based on time and semantic relatedness. In: *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Jaipur, pp. 1390–1396 (2016)
17. Gaurav, S., Jithendranath, Y., Adil, A., Yadav, S., Kasturi, B.: A study to assess and enhance educational specific search on web for school children. In: *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing, IEEE ICSC 2015*, pp. 260–263. Institute of Electrical and Electronics Engineers Inc. (2015)

Performance Comparison of Apache Spark and Hadoop Based Large Scale Content Based Recommender System

Saravanan S. ^(✉), Karthick K.E., Ashwin Balaji, and Anand Sajith

Department of Computer Science and Engineering, Amrita University, Bengaluru, India
s_saravanan@blr.amrita.edu, karthickke007@gmail.com,
ashwin.achu.1311@gmail.com, anandsajith@gmail.com

Abstract. The recommendation of products of interest to the user is pivotal for improving a customer's shopping experience. Recommender system has diversified and endeared itself in wide ranging industrial applications from e-commerce to online video sites. As the input data that is supplied to the recommender systems is large, the recommender system is often considered as data intensive application. In this paper, we present improvised MapReduce based data preprocessing and content based recommendation algorithms. Also, Spark based content based recommendation algorithm is developed and compared with Hadoop based content based recommendation algorithm. Our experimental results on Amazon co-purchasing network meta data show that Spark based content based recommendation algorithm is faster than Hadoop based content based recommendation algorithm. Also, graphical user interface is developed to interact with the recommender system.

Keywords: Hadoop · Spark · Recommender system

1 Introduction

A recommender system is a tool for the analysis of a large dataset and providing products that would be of interest to the user. The recommender system makes it easy for users to choose products which would be relevant to their tastes. The three basic approaches towards recommender systems are Content based recommender system, Collaborative filtering and hybrid recommender systems [1]. Content based recommender system is based on the notion that a user will be keen to buy a product that would be in the same category as the products in his inventory. Whereas, Collaborative filtering tries to find the similarity between users and the products bought by them, then it tries to forecast or predict the products-based on the products brought by similar users. User based collaborative filtering and item based collaborative filtering are the two basic types collaborative recommendation systems [2]. User based uses similarity between users to generate recommendations, while item based uses item ratings to generate similarity between products and recommends it to the customers. Recommender system analyzes large datasets to get accurate results. As it analyzes large data sets recommender systems are considered as data intensive application [3]. In recent times, Apache Hadoop and Apache Spark are considered to suit well for data intensive applications [4]. Remainder

of the paper is organized as follows. Section 2 represents the existing content based recommendation in Hadoop and Spark frame work and Sect. 3 explains the implementation of improved content based recommendation in Hadoop and content based recommendation in Spark and Sect. 4 presents the experimental setup and results. Section 5 presents the performance evaluation. Section 6 ends with the conclusion and future enhancements.

2 Related Work

De Pessemier et al. [5] developed content based Recommendation Algorithms on Hadoop for Wikipedia articles. They proposed MapReduce code for providing recommendations for the end users. Leskovec et al. [6] explain content based recommendation has to analyze massive dataset to provide good recommendations. Dooms et al. [7] propose In-memory, distributed content-based recommendation system which uses MapReduce paradigm. Generally, MapReduce parallel programming model stores mid-computation values in hard disk which is one of the drawbacks of it. In this paper, they proposed content based recommendation algorithm which keeps mid-computation values completely in RAM to reduce the hard disk accesses and to improve the efficiency of Map Reduce parallel programming model. Saravanan [8] developed large scale content based recommender system using Hadoop MapReduce Framework to provide N recommendations to the user and best recommendation to the user. The data preprocessing algorithm developed in [8] takes more time to execute. This is because data is not properly partitioned among the map tasks. So, in this paper we have overcome the drawback of the content based recommendation developed in [8]. We also have implemented the content based recommendation in Spark framework and evaluated the performance of Hadoop and Spark based content based recommender system. In this paper we have also designed the user interface to interact with the system.

3 Implementation

In this section we will present the improved Hadoop based content based recommender system and Spark based recommender system.

3.1 Data Set

The dataset [9] used in this paper is Amazon dataset which was collected in 2006. It contains the product metadata (Fig. 1).

<p>Id: Product id (number 0, ..., 548551)</p> <p>ASIN: Amazon Standard Identification Number</p> <p>Title: Name of the product</p> <p>Group: Group to which product belongs (Book, DVD, Video or Music)</p> <p>Salesrank: Amazon Salesrank (1 is highest rank and big numbers are lowest rank)</p> <p>Similar: ASINs of co-purchased products (i.e.) list of products similar to this product</p> <p>Categories: Location in product category hierarchy to which the product belongs</p> <p>Reviews: Product review information: time, customer id, rating, total number of votes on the review, total number of helpfulness votes</p>
--

Fig. 1. Amazon data set format

From the dataset we have observed the statistics listed in Table 1.

Table 1. Statistics table

No. of products	5,48,552
No. of customers	14,58,417
Customer bought most number of products	Customer whose ID is ATVPDKIKX0DER bought 109993 products

3.2 Improved Dataset Preprocessing Algorithm for Hadoop Based Content Based Recommendation Algorithm

MapReduce is one of parallel programming models, used for processing large amounts of data in parallel and distributed manner. Google company has introduced MapReduce programming model in 2004 [10]. This programming model is employed in Apache Hadoop framework [11]. The input data set is not properly partitioned among the map tasks in MapReduce based data set preprocessing algorithm proposed in [8]. In Hadoop 1, when you split the data set among map tasks, by default, the size of each split is equal to the size of block in HDFS which is 64 MB. So, every map task should get 64 MB for processing. But the customized record reader developed in [8] does not split the dataset properly among map tasks. In [8], each map task processes the whole data set as the input split implementation of customized record reader is not done properly. So, we overcome this drawback by modifying the customized record reader to partition data set properly among the map tasks. In our modified implementation, each map task will process only 64 MB data. Hence, the execution time of dataset preprocessing phase is less compared to dataset preprocessing algorithm proposed in [8].

3.3 Spark Based Content Based Recommendation Algorithm

This section presents implementation of Large Scale Content-based Recommender System in Spark. Apache Spark is a modern big data analytics framework which is based on the idea of Resilient Distributed Datasets which is an idea first published in [12] by a team of developers from AMPLab at the University of California at Berkeley. Spark is designed to overcome the drawbacks of MapReduce programming model. Spark uses RAM to store intermediate results whereas Hadoop uses hard disk to store intermediate results.

In Apache Spark, RDD is a collection of data items that are distributed across many compute nodes that can be processed in parallel. All our computations are expressed through operations on RDD that are automatically paralleled across the cluster.

Total Implementation is carried out in two steps. We implemented the following steps in Scala programming language [13].

Step 1: Find a list of products bought by every customer. This step is carried out in two sub steps.

Step 1.1: Scan the input file for every product. As we scan the file for every product, for every customer who has bought this product, we print the following information.
<Customer id: product id, similar items>

Step 1.2: The output produced in the previous step is grouped based on the customer id. The output of this step will contain the following information.

<Customer id: list of products with similar items for every product>

Step 2: Recommend products for each customer. This step is carried out in three sub steps.

Step 2.1: Scan the grouped data for every customer.

Step 2.2: Create a list of products the customer has bought and take two similar products for each product the customer has purchased and checks if the customer has bought those similar products.

Step 2.3: Remove the products which customer has already purchased from the list and print N products from the remaining items in the list. This N value will be passed as an input by the customer.

4 Experimental Setup and Results

4.1 System Configuration

We executed both improvised Hadoop based content based recommendation and Spark based content based recommendation algorithms on the system with the following configuration. The processor of the system is Intel Core i7 and its speed is 3.40 GHz. RAM capacity of the system is 8 GB. Ubuntu 16.04 is installed in the system. Our experiments are done with a cluster of size one node because the size of dataset is only 977 MB. If the data set size is huge then a cluster can be formed and used for analyzing the data. The version of Hadoop used is Hadoop 1.2.1. Spark is installed in the system. The dataset used for our experiment is Amazon dataset which was downloaded from Stanford University website [9].

4.2 Results

Graphical user interface of the recommender system is shown in Figs. 2, 3 and 4. The interface has the options for content based recommendation in Hadoop and content based recommendation in Spark. Figure 3 shows the recommendations of Hadoop framework. Figure 4 shows the recommendations of Spark Framework.

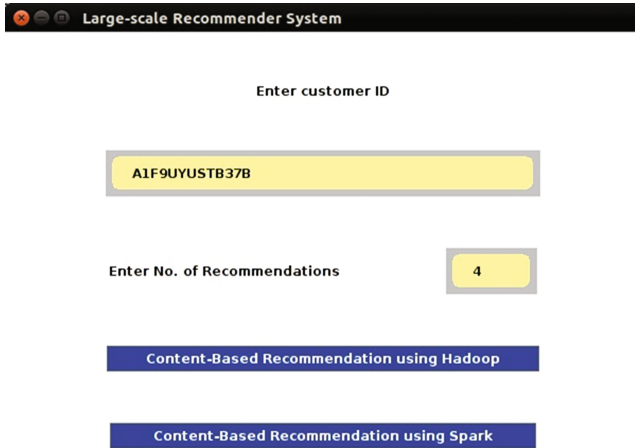


Fig. 2. Initial graphical user interface



Fig. 3. Output of content based recommendation in Hadoop

In interface from Fig. 2, the user of the recommender system has to enter the customer id for whom recommendations to be generated. After entering the customer id, user can mention how many recommendations should be generated. Then, user can choose either Hadoop based recommendation or Spark based recommendation for generating recommendations.



Fig. 4. Output of content based recommendation in Spark

Figure 3 shows the list of recommendations generated for the customer A1F9UYUSTB37B by Hadoop framework. There are 4 recommendations in the list as the user has chosen 4 as an input for number of recommendations.

5 Performance Evaluation

The Table 2 shows the time taken by improvised Hadoop and Spark based content based recommender system implemented in this paper.

Table 2. Execution time

Hadoop based Content Based Recommendation	Spark based Content Based Recommendation
Dataset preprocessing - 4.29 min	Find a list of products bought by every customer - 50 s
Content Based recommendation - 9 min	Content Based recommendation - 6.41 min
Total: 13 min 29 s	Total: 7 min 31 s

The graph in Fig. 5 pictorially shows the running time performance comparison of Apache Hadoop and Spark based content based recommender system. In y-axis, time is given in minutes. It is evident from the Table 2 and Fig. 5 that Spark based Content based recommender system generates recommendations faster than Hadoop based Content based recommender system.

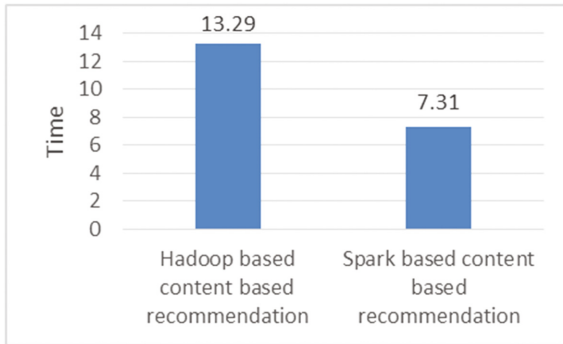


Fig. 5. Performance analysis of Hadoop based and spark based content based recommender system

6 Conclusion and Future Enhancements

We have improvised data set preprocessing algorithm for Hadoop based content based recommendation and developed Spark based Content based recommendation algorithm. The experimental results show that the Spark based content based recommendation algorithm generates recommendations faster than Hadoop based content based recommendation. This is because, in Hadoop, the intermediate results are stored in hard disk but in Spark the intermediate results are stored in RAM. So, as a future work, Spark based collaborative filtering recommendation system can be developed using an efficient clustering algorithm to group customers with similar interests to provide better recommendations.

References

1. Venkataraman, D., Gangothi, V., Saranya, S.: A comprehensive review of recommender system. *Int. J. Appl. Eng. Res.* **10**, 13909–13919 (2015)
2. Thangavel, S.K., Thampici, N.S., Johnpaul, C.I.: Performance analysis of various recommendation algorithms using apache hadoop and mahout. *Int. J. Sci. Eng. Res* **4**(12), 279–287 (2013)
3. Philip Chen, C.L., Zhang, C.-Y.: Data-intensive applications, challenges, techniques and technologies: a survey on big data. *Information Sciences*. Elsevier, Amsterdam (2014)
4. Kang, S.J., Lee, S.Y., Lee, K.M.: Performance comparison of OpenMP, MPI, and MapReduce in practical problems. *Adv. Multimed. J.*, Article ID: 575687. Hindawi Publishing Corporation (2014)
5. De Pessemier, T., Vanhecke, K., Dooms, S., Martens, L.: Content-based recommendation algorithms on the hadoop mapreduce framework. In: 7th International Conference on Web Information Systems and Technologies, pp. 237–240 (2011)
6. Leskovec, J., Rajaraman, A., Ullman, J.D.: Mining of Massive Datasets, pp. 322–331 (2014)
7. Dooms, S., Audenaert, P., Fostier, J., De Pessemier, T., Marten, L.: In-memory, distributed content-based recommender system. *J. Intell. Syst.* **42**(3), 645–669 (2014)

8. Saravanan, S.: Design of large scale content based recommender system using Hadoop MapReduce Framework. In: 2015 Eighth International Conference on Contemporary Computing (IC3). IEEE, 22 August 2015
9. <https://snap.stanford.edu/data/web-Amazon.html>
10. Dean, J., Ghemawat, S.: Mapreduce: simplified data processing on large clusters. *Commun. ACM* **51**(1), 107–113 (2008)
11. Apache-Hadoop. <http://Hadoop.apache.org>
12. Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., Franklin, M.J., Shenker, S., Stoica, I.: Resilient distributed datasets: a fault-tolerant abstraction for in memory cluster computing. In: Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation, p. 2. USENIX Association (2012). <http://www.cs.berkeley.edu/~matei/papers/2012/nsdispark.pdf>
13. Scala programming language. <http://www.scala-lang.org>

Performance Evaluation of AODV Routing Protocol for Free Space Optical Mobile Ad-Hoc Networks

Salma Fauzia^(✉) and Kaleem Fatima

Muffakham Jah College of Engineering and Technology, Osmania University,
Hyderabad, India
salmafauzial983@gmail.com

Abstract. RF based communication technologies have been the crux of wireless networking. The limitation of radio frequency (RF) based network is that the throughput deteriorates as new nodes are added to the network. Free space optics (FSO) technology is similar to fiber optics sans the fiber. It provides very high security and very high data rates. Free space optics (FSO) systems represent one of the most promising approaches for last mile connectivity, as compared to the other alternatives of fiber-optic cables, wireless local loops, and copper-based technologies.

A reactive protocol keeps in view the network dynamics during the process of routing and AODV routing protocol i.e. Ad-Hoc on demand distance vector is designed for wireless Ad-Hoc networks. It derives the advantages of both Destination sequenced distance vector routing and Dynamic source routing. In this work we run simulations of AODV under a RF based and FSO based MANET scenario and support our conclusions that AODV when modified to handle multiple interfaces results in throughput improvement by 5X, and achieves packet delivery ratio of 96%.

Keywords: AODV · FSO · RF · Nodes

1 Introduction

As the data communication networks become denser, it gives rise to problems related to capacity and also limited unlicensed part of the electromagnetic spectrum paves way for exploring the optical part of the electromagnetic spectrum. Free space optical communication is characterized by high bandwidth, license free band of operation, spatial reusability. Merger of this technology with Mobile Ad-Hoc networking capabilities is called as a Free Space Optical Mobile Ad-Hoc Network. Yuksel et al. [13] introduced the basic building blocks for these type of networks and their work deals with issues related to routing and localization. Free space optical MANETs face the challenge of maintaining LOS i.e. line of sight between the nodes and also weather conditions affecting it very badly. Yuksel et al. [12] developed solutions for these issues by arranging the transceivers spherically on a node and also by using multihop type of communication.

2 Related Work

Proactive protocols like DSDV [1] and OLSR [11] regularly broadcast routing information across the network or in certain areas of the network, and maintain robust data structures called routing tables at each node. Protocols like AODV [2] and DSR [3] are categorized under reactive protocols that perform route discovery by broadcasting i.e. flooding the network. Until a route is found, there is latency in data forwarding. Flooding causes wastage of network resources and as networks increase in size and complexity, new methods to limit flooding were necessary. Hierarchical routing protocols such as HRP [4], LANDMAR [6] entwine the network into areas that maintain routing information within the area. Within each coverage area a node selected as gateway node maintains routing tables that interact with the other gateway nodes. Thus, routing within each coverage area happens normally while routing in inter-coverage area is handled by the gateway node. Though this technique helps in dealing with scalability issues, and proves to be an important step in achieving greater scalability, the increased intricacy of reformation makes it harder to implement these routing techniques as they lean on the gateway nodes that maintain routing between regions. Chances of failing at a single point due to failure of a gateway node are more. When we use directional antennas the existing routing protocols fail to perform satisfactorily. It is very interesting to study how these protocols behave when modified to handle directional antennas. To address issues with interface handoff, backoff, and neighbor discovery, Choudhury et al. [5] proposed Directional DSR (DDSR), a modification to DSR [3] and a cross layer protocol inspired by DSR which handles route discovery, establishment, maintenance, and route recovery mechanisms using directional antennas was studied by Gossain et al. [8] that presents Directional Routing Protocol (DRP). While much of these efforts in using directionality in the routing layer are important, they come more as a response to having directional communications rather than leveraging directionality as a benefit in routing. Very limited work shows how these protocols can be adapted for a free space optical MANET scenario. Keeping the directional nature of FSO transceivers in mind, Nasipuri et al. [7] modified the RTS and CTS exchange in 802.11 to support directionality and showed through simulations a throughput improvement of 2–3 times over omnidirectional antennas. The notion of directionality at layer 3 was studied by Murat Yuksel et al. [9] as they proposed Mobile orthogonal rendezvous routing protocol, and also lead to conclusions how these type of protocols can be adapted for a FSO MANET. In this work we compare the performance of AODV routing protocol in RF based and FSO based MANET and show how throughput and packet delivery fraction are affected with varying number of nodes and mobility. Section 3 presents the methodology of AODV routing protocol for RF and FSO based MANET. Section 4 presents graphical results of simulations. Section 5 deals with conclusion and future scope.

3 Protocol Basic Methodology

The basic protocol AODV [2] generates routes to the destination on demand. The protocol selects the shortest path to the destination node. It is specially designed to handle mobility and it takes care of changes in topology and repairs routes in case when

the node moves in and outside the communication range. HELLO messages are sent out at regular intervals so that a node keeps knowledge of its neighbor. Every node has a unique ID called as the sequence number. During the process of discovering the route to the destination, a node sends out a request message typically called as RREQ. The sequence number along with destination node address is a part of RREQ packet. If the node receiving the request packet has the path to the destination, it sends out a reply message called as RREP to the node which generated the RREQ, else sends out the request packets again i.e. rebroadcasts the request packet. To have knowledge of the recent route, sequence numbers are used. Higher sequence number indicates the most freshest route. The dynamic nature of the network causes the nodes to move in and move out of the communication range, hence route error messages RERR are generated to tackle this issue. When a route error message is received by a node it removes the entries of all the nodes which cannot support routing from the routing table. An FSO transmitter and receiver pair should be aligned properly to overcome LOS, so that these request and reply packets are received by the nodes promptly. A sample network with six mobile nodes is considered for explanation as shown in the Fig. 1. The “Req” messages initiated by source S i.e. N1 is broadcast. These messages are forwarded by the nodes till it reaches the destination D i.e. N6. To depict realistic scenario, we simulated a network of 100 mobile nodes on NS-2.34. The pause time i.e. the time for which the node rests at a place before it starts moving again is 0 s and a random way point model is used. The links represent the nodes that can communicate with each other and not a wired link.

P1, P2, P3 depict the various paths from source to destination.

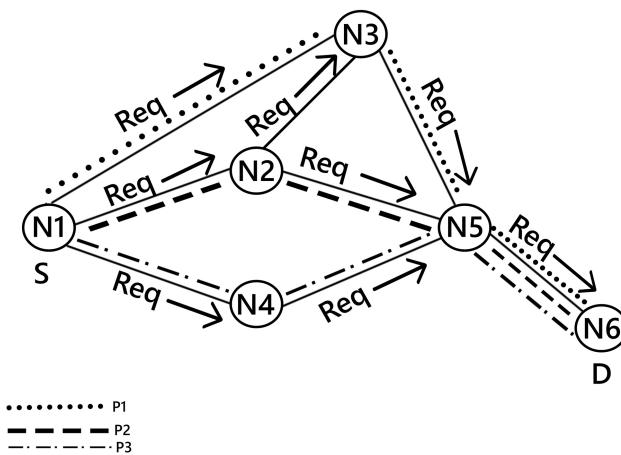


Fig. 1. Route request propagation showing multiple choice of paths from source to destination

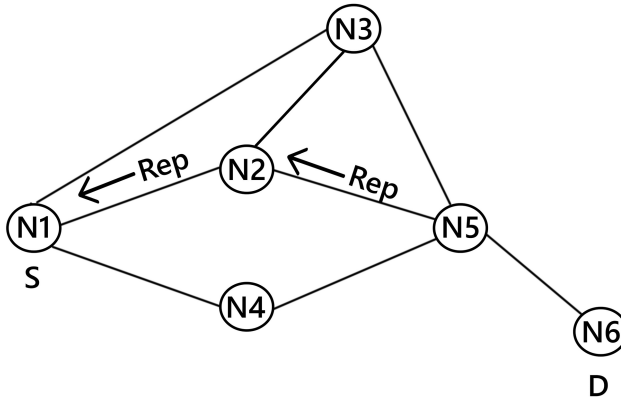


Fig. 2. Example of a route reply from N5-N2-N1

Node N5 and N2 have a route to the destination and hence send a “Rep” message i.e. reply message is propagated to the source node S as shown in Fig. 2.

Figure 3 is a partial trace format obtained after simulation of 100 mobile nodes which shows the request messages for nodes numbered 94,38,97,61, and reply messages for node numbered 23.

```

r 0.557979781 94 RTR --- 0 AODV 48 [0 ffffffff 2f 800] ----- [47:255 -1:255 22 0] [0x2 9 1 [12 0] [3 4]] (REQUEST)
r 0.557979938 38 RTR --- 0 AODV 48 [0 ffffffff 2f 800] ----- [47:255 -1:255 22 0] [0x2 9 1 [12 0] [3 4]] (REQUEST)
r 0.559677590 23 RTR --- 0 AODV 44 [13a 17 8 800] ----- [12:255 3:255 29 23] [0x4 2 [12 4] 10.000000] (REPLY)
f 0.559677590 23 RTR --- 0 AODV 44 [13a 17 8 800] ----- [12:255 3:255 28 91] [0x4 3 [12 4] 10.000000] (REPLY)
s 0.561791621 97 RTR --- 0 AODV 48 [0 ffffffff 3d 800] ----- [97:255 -1:255 20 0] [0x2 11 1 [12 0] [3 4]] (REQUEST)
r 0.563079928 61 RTR --- 0 AODV 48 [0 ffffffff 61 800] ----- [97:255 -1:255 20 0] [0x2 11 1 [12 0] [3 4]] (REQUEST)
    
```

Fig. 3. Trace format (partial) depicting request and reply packets being forwarded and received for a 100 node scenario

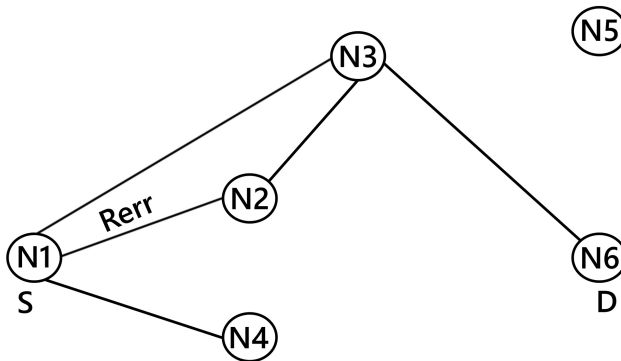


Fig. 4. Example of a route error propagation, N5 moved out of range

```

s 0.893726228 _84_ RTR --- 0 AODV 32 [0 0 0 0] ----- [84:255 -1:255 1 0] [0x8 1 [12 0] 0.000000] (ERROR)
r 0.894546385 _7_ RTR --- 0 AODV 32 [0 ffffffff 54 800] ----- [84:255 -1:255 1 0] [0x8 1 [12 0] 0.000000] (ERROR)
r 0.894546512 _63_ RTR --- 0 AODV 32 [0 ffffffff 54 800] ----- [84:255 -1:255 1 0] [0x8 1 [12 0] 0.000000] (ERROR)
    
```

Fig. 5. Trace format(partial) depicting route error messages for a 100 node scenario

Figure 4 depicts route error messages “Rerr”, when node N5 moves out of range. Figure 5 shows the error messages sent back in an event when there is a loss in connection due to node moving out of range. Here, nodes numbered 84,7,63 are unable to participate during the process of routing.

Bilgi and Yuksel [10], in their study introduced the various modules to be used along with NS-2.34 for an FSO MANET. The node will have a separate stack from Physical layer up to Link layer for each transceiver. This is attached to the AODV routing agent as shown in Fig. 6. These modifications were done to handle transmission over multiple interfaces. The angle and position of the node is taken care by the channel and this information is obtained from the packet headers. The packets are delivered to the upper layers and this action is “planned” by the channel for the transceiver. When the position of the node changes, the auto alignment circuitry is responsible for changing from one interface to another [14].

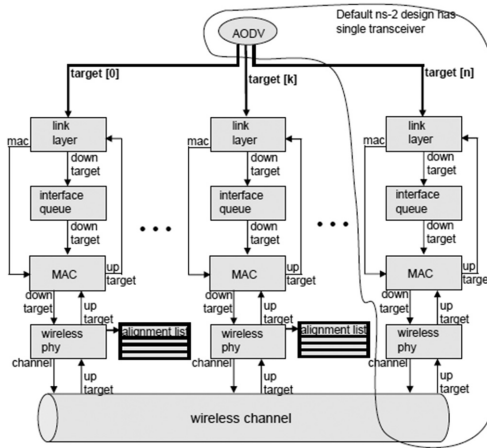


Fig. 6. FSO node structure in NS-2 [14]

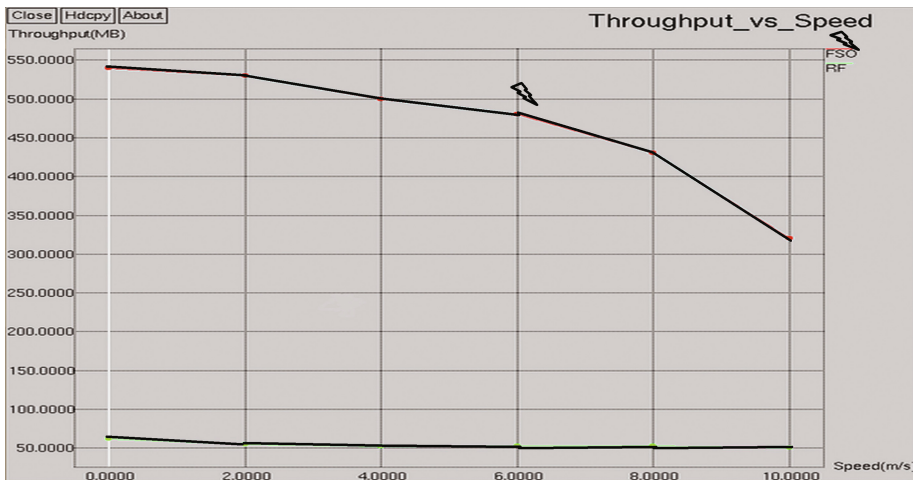
4 Simulation Results and Observations

In this section the simulation parameters are listed and a brief explanation about the graphs is given. The nodes were considered to be spread over an area of 1300 m*1300 m. The FSO-extension package was merged with NS 2.34 for FSO simulations. The unaltered version was used to simulate AODV in RF (Table 1).

Table 1. Simulation parameters

Parameter	Value
Number of nodes	50,100
Simulation time	500 s, 70 s
Network range	1300 m * 1300 m
Transmission range	250 m
Traffic type	CBR
Packet size	512 bytes
Maximum speed	30 m/s
Simulator	NS-2.34, FSO-extension package

The speed of the mobile nodes was varied from 2 m/s to 10 m/s and the throughput was observed to deteriorate with increasing speed. The node mobility causes loss in connectivity and handling the same for multiple interfaces as node moves out of range becomes difficult. The throughput was maximum 550 MB when the node mobility was 0 m/s. As observed from Fig. 7 throughput is higher for FSO MANET and this can be credited to directional transmission over interfaces. By keeping the area constant when the number of nodes are increased from 20 to 120 with a node speed of 4 m/s Fig. 8 shows there is a fall in throughput which is more severe for RF MANET because of its omnidirectional nature of transmission.

**Fig. 7.** Throughput of the network as a function of speed of mobile nodes (m/s)

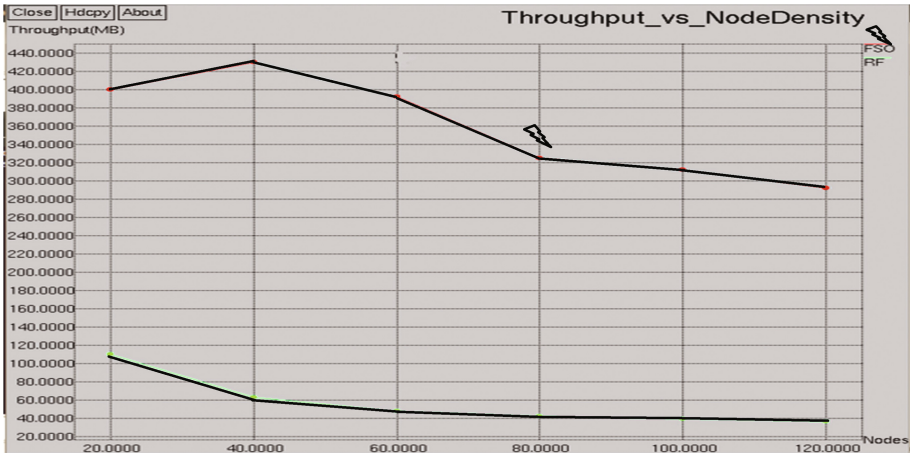


Fig. 8. Throughput as a function of increasing number of nodes in the network

Figure 9 illustrates when the source destination connections are increased the throughput of FSO MANET is better than RF MANET, because the network is inundated with request and reply packets which affects the throughput of the RF network, whereas LOS alignment and spatial reuse improve the throughput of FSO MANET considerably. This simulation is performed for 50 nodes for a time period of 500 s. FSO MANET is successful in delivering the packets to the destination but decreases gradually because of non alignment of directional antennas with increasing source destination pairs as shown in Fig. 10.

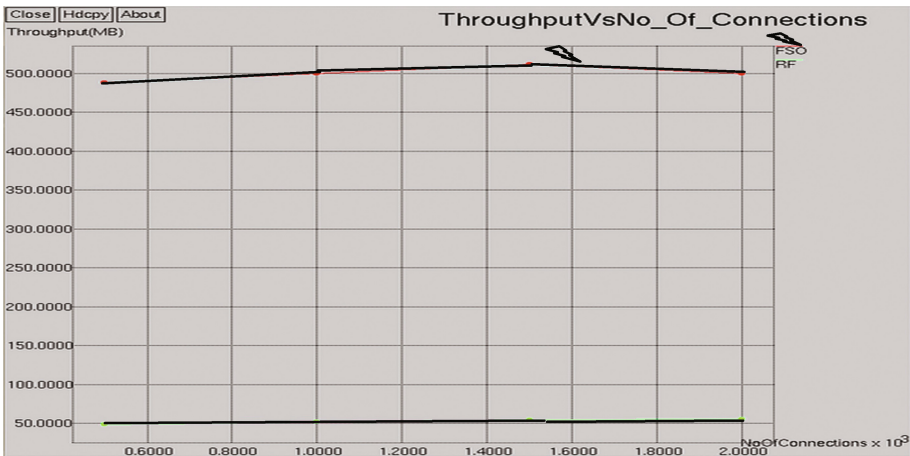


Fig. 9. Throughput as a function of number of connections between source to destination

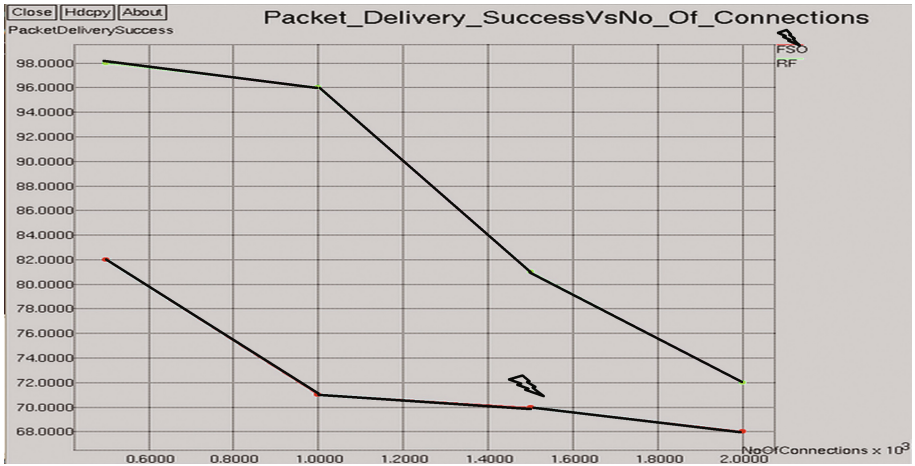


Fig. 10. Packet delivery success as a function of number of connections between source to destination

It is rather fascinating to observe the time delay in receiving the packets at the destination. Packet delivery success for a 100 node scenario with speeds varied in steps of 10 m/s, 20 m/s, 30 m/s as depicted in Fig. 11 and Fig. 12. Transmission over interface reduces meddling of signals. AODV displays a broadcast nature and hence with increasing speeds the successful delivery of packets to the destination is affected as more number of retrial attempts storm the network.

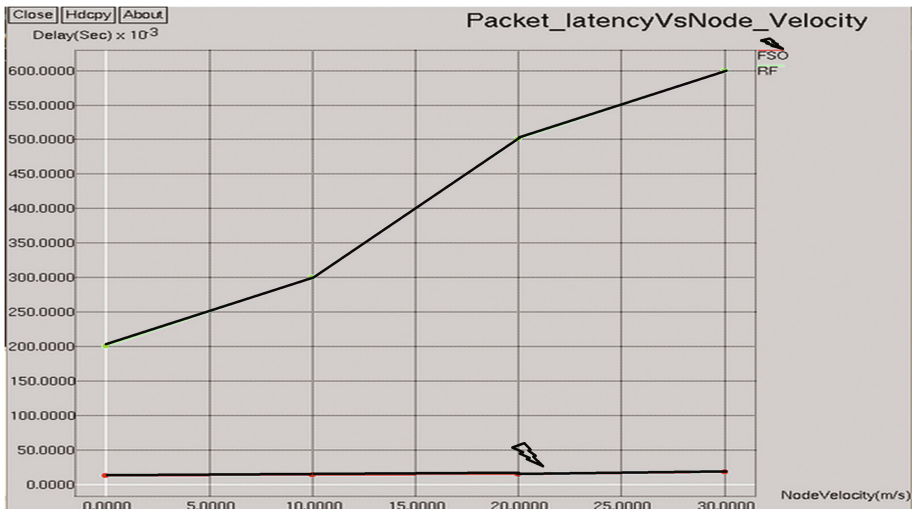


Fig. 11. Data packet latency as a function of node velocity (m/s), 100 nodes, 1300*1300 area

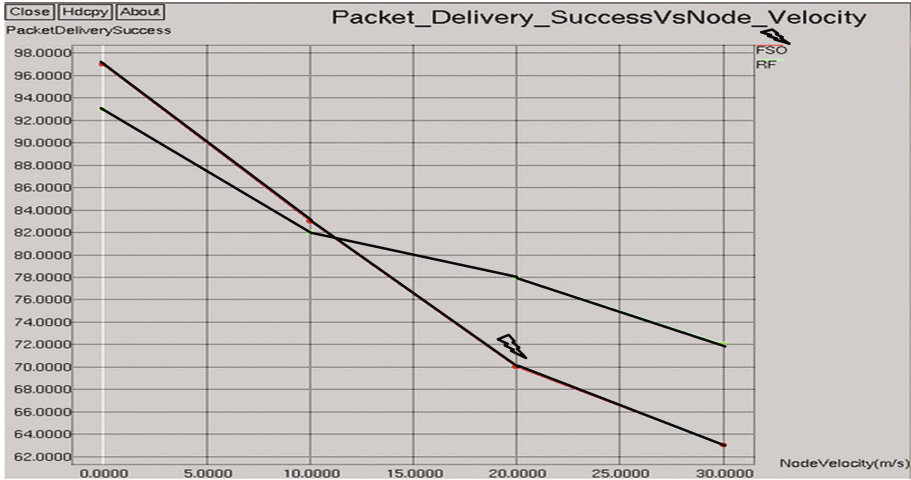


Fig. 12. Packet delivery success as a function of node velocity (m/s), 100 nodes, 1300 m*1300 m area

5 Conclusion

Sheltered communication and low cost equipment add to the benefits of using the optical part of the EM spectrum. Though FSO technology is heavily affected by the environmental conditions, it can act as an excellent back up for RF based networks. Present nanometer technology can yield designs that can lead to many more improvements at the Physical layer which is responsible for maintaining LOS [14].

Future work can be extended to various domains in routing, localization and positioning mechanisms for the benefit of providing quality of service to its users.

References

1. Perkins, C., Bhagwat, P.: Highly dynamic destination-sequenced distance-vector routing (DSDV) for mobile computers. In: Proceedings of ACM SIGCOMM, pp. 234–244 (1994)
2. Perkins, C., Royer, E.M.: Ad hoc on-demand distance vector routing. In: Proceedings of the 2nd IEEE Workshop on Mobile Computing Systems and Applications, New Orleans, LA, pp. 90–100, February 1999
3. Johnson, D., Maltz, D., Broch, J.: DSR: the dynamic source routing protocol for multi-hop wireless ad hoc networks, Chapter 5. In: Perkins, C.E. (ed.) Ad Hoc Networking, pp. 139–172. Addison-Wesley, Boston (2001)
4. Tsai, W.T., Ramamoorthy, C.V., Tsai, W.K., Nishiguchi, O.: An adaptive hierarchical routing protocol. IEEE Trans. Comput. **38**(8), 1059–1075 (1989)
5. Choudhury, R.R., Vaidya, N.: Impact of directional antennas on ad hoc routing. In: Proceedings of the Eighth International Conference on Personal Wireless Communication (PWC), Venice, September 2003

6. Gerla, M., Hong, X., Pei, G.: Landmark routing for large ad hoc wireless networks. In: Proceedings of IEEE GLOBECOM, San Francisco, CA, Nov 2000
7. Nasipuri, A., Ye, S., Hiromoto, R.E.: A MAC protocol for mobile ad hoc networks using directional antennas. In: Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC 2000), Sept 2000
8. Gossain, H., Joshi, T., De Morais Cordeiro, C., Agrawal, D.P.: DRP: an efficient directional routing protocol for mobile ad hoc networks. *IEEE Trans. Parallel Distrib. Syst.* **17**(12), 1439–1451 (2006)
9. Cheng, B., Yuksel, M., Kalyanaraman, S.: Using directionality in mobile routing (short paper). In: Proceedings of IEEE International Conference on Mobile Ad-hoc and Sensor Systems (MASS), Atlanta, GA, Sept 2008, pp. 371–376
10. Bilgi, M., Yuksel, M.: Packet-based simulation for optical wireless communication. In: Proceedings of IEEE Workshop on Local and Metropolitan Area Networks (LANMAN), Long Branch, NJ, pp. 1–6, May 2010
11. Clausen, T., Jacquet, P.: OLSR RFC3626, October 2003. <http://ietf.org/rfc/rfc3626.txt>
12. Yuksel, M., Akella, J., Kalyanaraman, S., Dutta, P.: Free-Space-Optical Mobile Ad-Hoc Networks: Auto-Configurable Building Blocks. *ACM/Springer Wireless Networks* (2007)
13. <http://www.ece.ucf.edu/~yukse/fso-MANET.htm>
14. Nakhkoob, B., Bilgi, M., Yuksel, M., Hella, M.: Multi-transceiver optical wireless spherical structures for MANETs. *IEEE J. Sel. Areas Commun.* **27**(9), 1612–1622 (2009)

Combination of Fuzzy Logic Digital Image Watermarking and Advanced Encryption Technique for Security and Authentication of Cheque Image

Sudhanshu Suhas Gonge^{1(✉)} and Ashok Ghatol²

¹ Faculty of Engineering and Technology,
Sant Gadge Baba Amravati University, Amravati, India
sudhanshu1984gonge@rediffmail.com

² Dr. Babasaheb Ambedkar Technological University,
Lonere, Maharashtra, India
vc_2005@rediffmail.com

Abstract. The demonetization and corruption can be stop by doing cashless payment in country with help of card and cheque transaction payments. However, on card transaction additional charges are applied whereas; there are no extra charges applied by bank to the customer while doing payments through cheques. The bank uses cheque truncation system for faster clearance of customer cheques. There are many methods and techniques used for providing the authorization service to digital image. In this research work, the digital image watermarking is used with the help of Fuzzy logic technique using dynamic fuzzy interference system. The security services are provided to watermarked cheque image using 256 bits key advanced encryption standard. This results the authorized and secured transmission of cheque document image. However, the performance and analysis of this research is done by applying various types of attacks.

Keywords: Digital watermark · DFIS · AES · Attacks · Fuzzy logic

1 Introduction

The corruption is one of the major drawbacks for developing country. In day-to-day life, small business can be done on the basis of cash payments. However, some of the business people are using cash for bribe. There are many things, which are cheaper in cost but it is sold in its double cost by accepting only cash. To stop bribery and corruption, it is necessary to make cashless payment. Many people avoid paying the income tax based on their annual income. It is also been observed minimum charges are applied by bank to customer, if payments are done through cards, demand drafts, NEFT, and RTGS [1]. There are also drawbacks by doing payments through cash mode. Since, some of the currency notes occurring in the daily routine business are pirated. To overcome this issue, only secure cheque payment can be used. It may help the country for making cashless and free from corruption. The cheque payments can be done using CTS of bank. In this process, the cheque document is scanned and then it is

send to the clearing house of the bank for transferring the amount from one account to another. To provide authorization to the digital image of bank cheque digital image watermarking is used. However, the digital watermarking process is carried out on the basis of various properties, methods and techniques. The digital watermarking is classified as shown in Fig. 1.

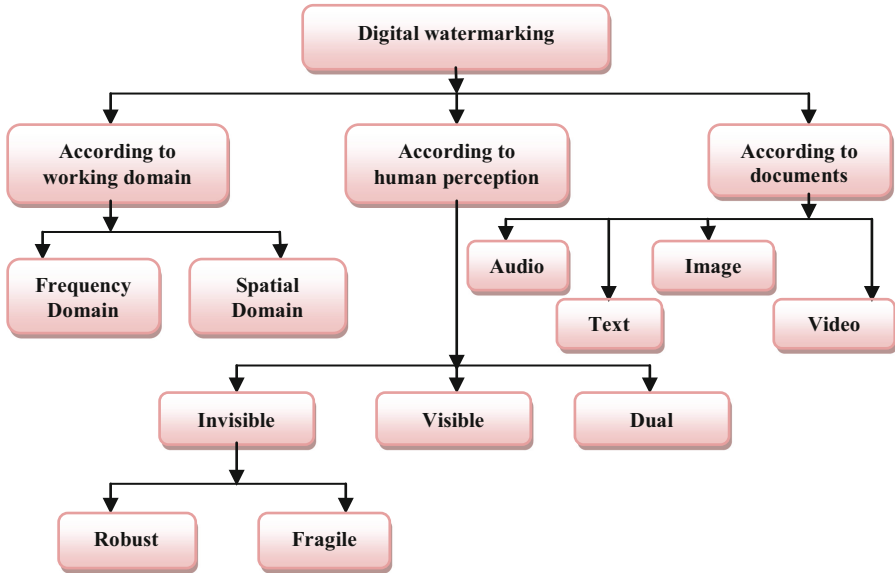


Fig. 1. Digital watermarking classification.

In this paper, “*Combination of Fuzzy Logic Technique Using Dynamic Fuzzy Interference System Used For Digital Image Watermarking along with AES technique using 256 bits key*” method is discussed for authorization and security of bank cheque image. The performance and evaluation of this method is done by applying various attacks viz; (i) Cropping, (ii) Gaussian Blur, (iii) JPEG Compression, (iv) Median Filter, (v) Rotation, (vi) Salt & Pepper Noise, and (vii) Under Normal Mode on combined watermarked and encrypted bank cheque image [2–12]. Many researchers has proposed digital image watermarking schemes using frequency domain techniques. To provide best scheme, the research is performed by purposing the hybrid combination of these two technique to enhanced the robustness of watermark quality and security of bank cheque image which is been explained in [13–17].

2 Techniques Used for Algorithms

There are two main algorithms i.e. (i) Watermarked embedding and encryption algorithm, and (ii) Decryption of watermarked image & watermark extraction algorithm [13]. To implement these algorithms, the three techniques are used viz. (i) Discrete wavelet transform, (ii) Fuzzy Logic, and (iii) Advanced encryption standard technique.

2.1 Discrete Wavelet Transform

The wavelet is a wave function of the signal, which can be obtained by applying sampling techniques on signal, and set of wavelet function is achieved [5–7]. There are many wavelet functions which are being derived from their mother wavelet function shown in Eq. 1 [4–7].

$$\psi_{x,y}(t) = \frac{1}{\sqrt{x}} \psi\left(\frac{t-y}{x}\right) \quad (1)$$

Where,

x is scaling factor and y is the shifting parameter of mother wavelet signal function.

There are different wavelets function like:-

- (i) Morlet wavelet,
- (ii) Daubechies wavelet,
- (iii) Continuous wavelet, and
- (iv) Haar wavelet, etc.

In this paper, 1-D Haar wavelet is used & applied on the image to decompose it into four non overlapping bands.

2.2 Fuzzy Logic

Fuzzy logic is a set of mathematical principles for knowledge representation based on degrees of membership. It is basically consist of multi-valued and deals with membership and degree of truth. Every fuzzy logic model uses the continuum of logical value occurs between zero (completely false) & 1 (completely true). It is a set of fuzzy boundaries. It is formally defined as a fuzzy set 'A' in X is represented as set of ordered pair. A fuzzy set totally characterized by membership function shown in Eq. 2 [8].

$$A = \{(x, U_A(x)) | x \in X\} \quad (2)$$

Fuzzy logic technique is based on types of rules like:-

- (i) Mamdani Assilian model,
- (ii) Takagi-Sugeno model,
- (iii) Classifier model, etc.

However, Dynamic fuzzy inference system using fuzzy rule based classifier is applied for watermarking purpose. The dynamic fuzzy inference system is also known as 'Expert System', which works on the base of rules. The working principle of dynamic fuzzy inference system (DFIS) is explained as shown in Fig. 2.

The dynamic fuzzy inference accepts the crisp input on which fuzzification is done using fuzzifier with the help of fuzzy classifier model rule, which results into fuzzy output. This fuzzy output is given fuzzy inference engine. The same set of fuzzy classifier model rules are used by fuzzy inference engine and defuzzification of data is done with the help of defuzzifier to generate the crisp output [9–12].

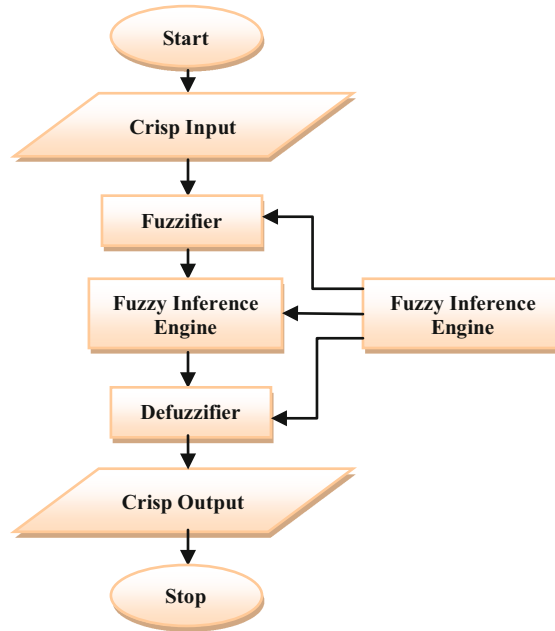


Fig. 2. Dynamic fuzzy inference system.

2.3 Advanced Encryption Standard

There are many encryption and decryption techniques used for providing the security to digital data. Some of the techniques are shown in following Table 1 based on there properties and methods like:-

- (1) Data Encryption Standard (DES),
- (2) Triple Data Encryption Standard,
- (3) International Data Encryption Algorithm (IDEA),
- (4) CAST-128,
- (5) RC4,
- (6) Advanced Encryption Standard (AES), and
- (7) Blowfish Encryption Algorithm.

These above algorithms performs the mathematical operations like addition, subtraction, XOR, fixed S-Boxes, Permutation and Substitutions using variables based on key size and block size of data. The following Table 1 shows the types of algorithms, size of key used for operations, number of rounds performed by algorithm and their applications [3, 6].

Table 1. Types of encryption algorithm and their applications.

Types of encryption algorithm	KEY	Number of rounds	Application
DES	56	16	SET, Kerberos
Tiple DES	112 or 168	48	PGP, S/MIME
IDEA	128	8	PGP
CAST-128	40 to 128	16	PGP
RC5	Variable to 2048	Variable to 255	Security to Databases
AES	128 or 192 or 256	10 12 14	Security to Sensitive Data
Blowfish	Variable to 448	16	Password Management

The AES technique is applied on the data based on its block size and key length. There are three keys generally used viz, (i) 128 bits key take 10 rounds, (ii) 192 bits key takes 12 rounds, and (iii) 256 bits key takes 16 rounds while execution of AES encryption and decryption operation [3,4,5, and 6]. The 256 bits key AES technique is applied on fuzzy watermarked cheque image to provide security service for bank cheque document. The working principle is shown in Fig. 3.

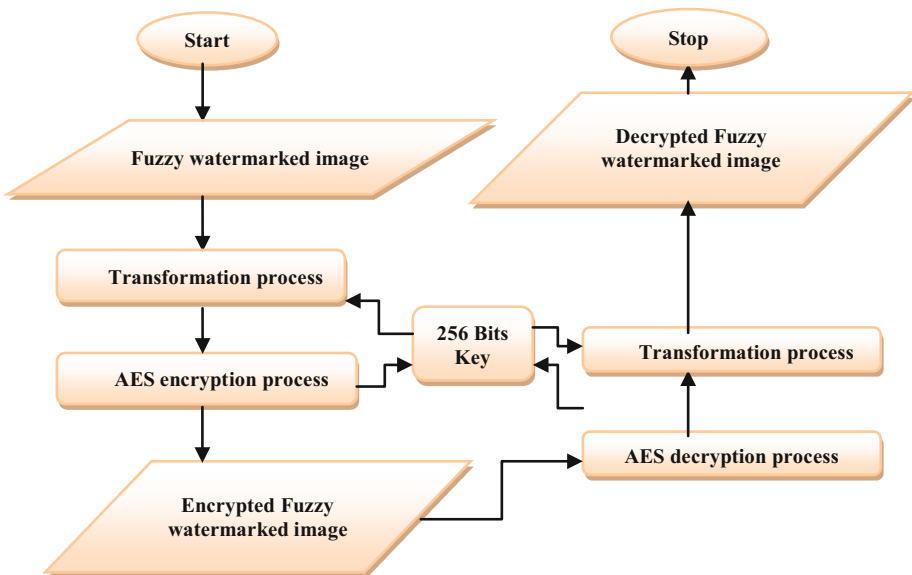


Fig. 3. Working of AES encryption and decryption process applied on fuzzy watermarked image.

3 Propose Algorithm

The proposed digital cheque watermarking technique is basically classified into 3 parts as shown in Fig. 4.

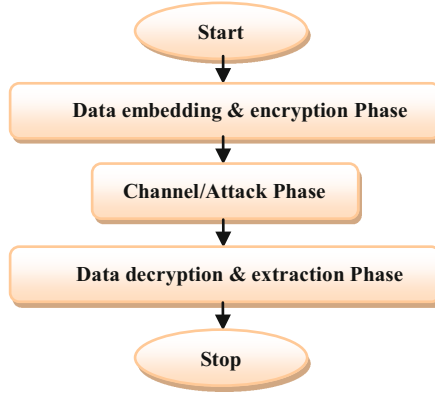


Fig. 4. Basic data flow diagram of proposed algorithm.

3.1 Watermark Embedding and Cheque Image Encryption Algorithm

The steps are as follows:-

- Select bank cheque image.
- Resize color image of cheque into 512×512 pixel and select blue plane for embedding watermark.
- Perform 1-level of Haar wavelet transform on selected blue plane to obtained approximate co-efficient of cheque image perform quantization operation.
- Select watermark logo image and generates a watermark formation in vectors of 0's & 1's.
- Create two PN_Sequences of zero's and one's from watermark formation which is exactly equal to same using gain factor $\beta = 0.5$.
- Texture sensitivity is calculated of selected components approximate band to embed watermark and apply these coefficient to DFIS.
- Apply fuzzy inference rule to DFIS to generate watermark weighting factor.
- Perform watermark embedding process in approximate co-efficient sub-band of DWT image using equation:-

$$I'_{j+\beta} = \text{DFIS} \left(\sum j * \text{round} (I_{j+\beta}/Q) \right) + X_j$$

Where,

Q is Quantization value.

$I'_{j+\beta}$ is co-efficient of watermarked image.

- Perform inverse of 1-level DWT Haar transform on approximate co-efficient band to get watermarked image.

- (j) Apply AES encryption technique on watermarked image using 256 bits key.
- (k) Finally, combination of fuzzy logic watermarked and AES encrypted bank cheque is obtained.

The work flow diagram for DFIS watermarking & AES encryption process of cheque image is shown in Fig. 5.

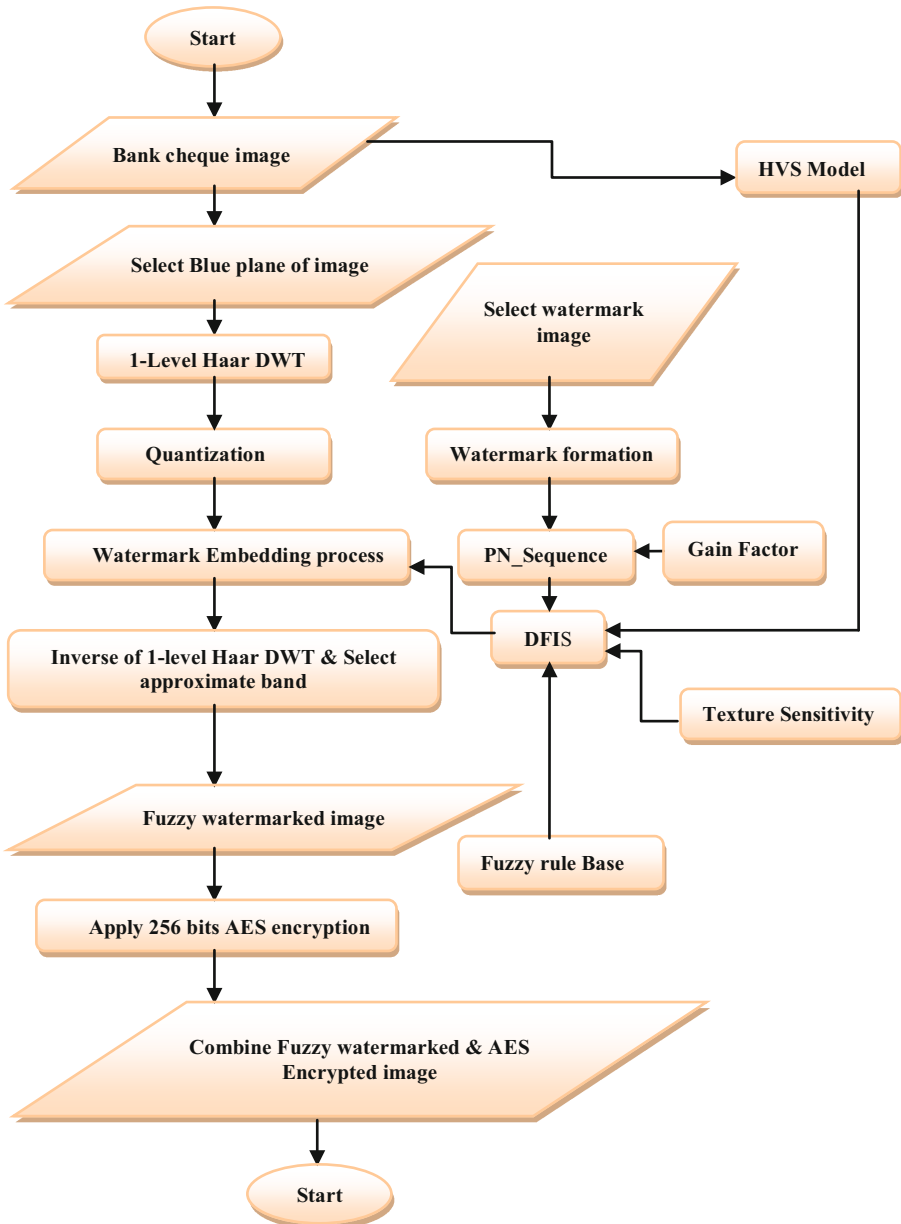


Fig. 5. DFIS watermarking and AES encryption process of cheque image.

3.2 Cheque Image Decryption and Watermark Extraction Algorithm

The Fig. 6 explains work flow diagram for cheque image decryption and watermark extraction process.

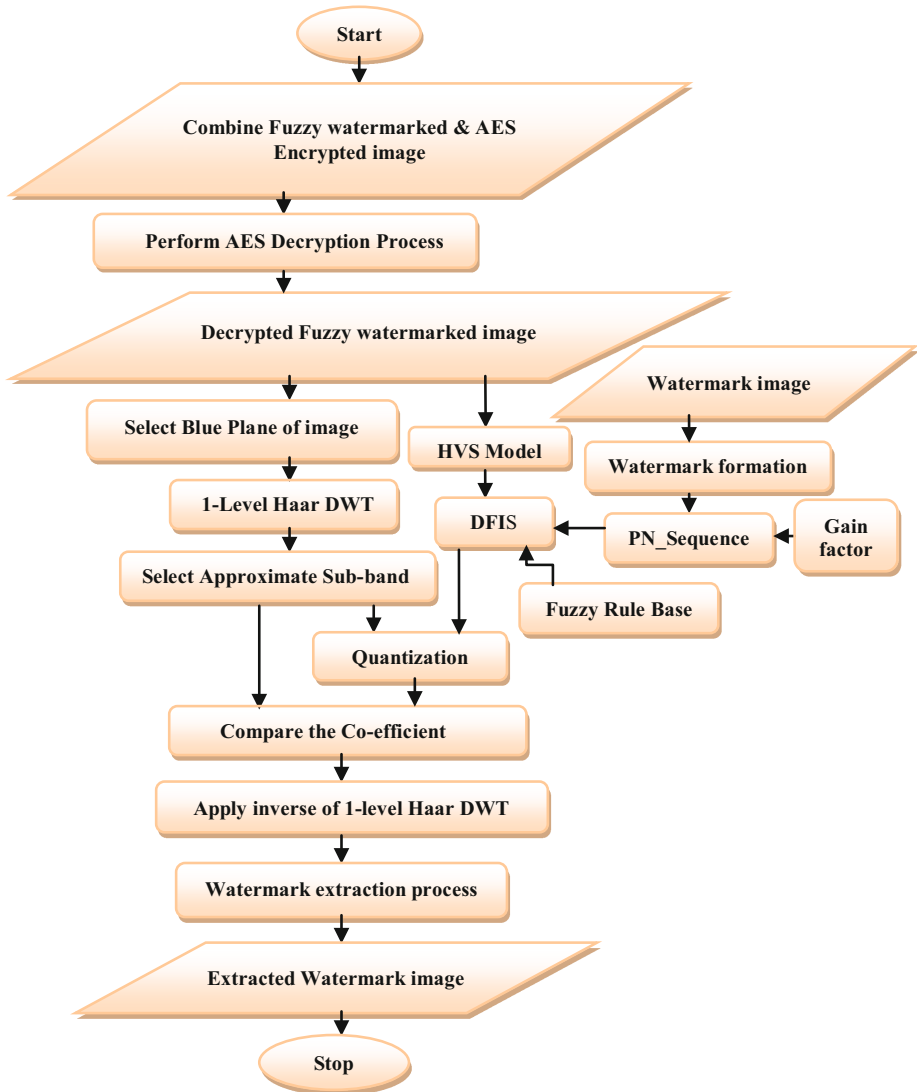


Fig. 6. AES decryption & watermark extraction process of cheque image using DFIS.

The steps are as follows:-

- (a) Select fuzzy watermarked & AES encrypted bank cheque image.
- (b) Perform 256 bits key AES decryption process.
- (c) Decrypted fuzzy watermarked image is obtained.
- (d) Select the blue plane of decrypted & fuzzy watermarked image.
- (e) Apply 1-level Haar wavelet transform.
- (f) Select approximate co-efficient of sub-band.
- (g) Perform quantization operation on approximate co-efficient of DWT $I''(j)$ by Q and apply to DFIS.
- (h) Extract the watermark using Eq. 4.

$$W'(j) = I''(j) - DFIS \left(\sum \text{round} (I''(j)/Q) \right) \tag{4}$$

- (i) Reconstruct the extracted watermark bits & calculate similarity between original watermark and extracted watermark.

4 Parameter Used for Evaluation of Performance

The performance of digital image watermarking is evaluated by using basic four parameter viz, (i) Peak signal to noise ratio, (ii) Mean square error, (iii) Robustness, and (iv) Time [5–17].

5 Results and Discussion

In this experiment, the cheque image of 512×512 sizes of JPEG format, having resolution of 96 dpi vertically and horizontally with a depth of 24 bits is selected as a host image shown in Fig. 7. The watermark image used for embedding also has same dimensions as that of the host cheque image. The watermark image is shown in Fig. 8.

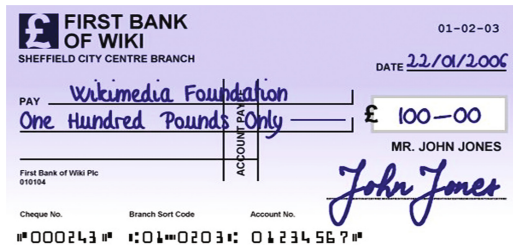


Fig. 7. Original cheque image.



Fig. 8. Watermark image.

These two images are used as basic images to perform digital watermarking process and AES encryption technique is applied using 256 bits key on watermarked cheque image. In this process, there are sub-process as explained in point 4.4. These process consume time for performing operation as explain in proposed algorithm.

Table 2. Time taken by embedding, encryption, decryption, and extraction process against various attacks.

Different types of Attacks	Watermark embedding time in seconds	Watermarked image encryption time in seconds	Watermark extraction time in seconds	Watermarked image decryption time in seconds	Complete elapsed time in seconds
Cropping	10.701	0.171	6.957	0.156	0.0803
Gaussian Blur (0.04 dB)	9.906	0.202	6.988	0.343	0.0942
JPEG Compression	9.843	0.202	7.332	0.202	0.0792
Median Filtering	9.999	0.171	7.300	0.280	0.0808
Rotation (45°)	9.999	0.171	7.004	0.202	0.0776
Salt & Pepper Noise	9.952	0.171	6.910	0.202	0.0811
Under Normal Mode Attack	10.608	0.202	7.160	0.171	0.0776

Table 2 shows the time taken by each process against various attacks explained in research work. The Fig. 9 explains the graphical representation of time versus against different attacks plotted from Table 2. In this experiments, it is observed that the time consume by embedding process is quite high as compared to that of other. From graph shown in Fig. 9, the embedding time taken by watermark in cheque image is same against rotation attack and median filtering attack. The embedding time of watermark against cropping is very high as compared to that of the rest of attacks used in experiment. From graph it is seen that the encryption time taken by AES technique using 256 bits key is 0.171 s against cropping, median, rotation, and salt & pepper noise attacks whereas, it is observed that 0.202 s are taken against JPEG compression, under normal mode & gaussian blur attack with an intensity of 0.04 dB. It is also

observed that extraction time taken by watermark extraction process is comparatively less as compared to that of watermark embedding process. The watermark extraction time against salt and pepper noise attack is 6.910 s, which is comparatively low as that of the remaining attacks used in this research. However, it is also observed that the time taken for decryption of watermarked cheque image is 0.202 s against JPEG compression, rotation and salt & pepper noise attack whereas 0.343 s against Gaussian blur attack which is comparatively high as that of other attacks. The minimum time taken for decrypting watermarked cheque image is 0.156 s against cropping attack. The complete elapsed time is 0.776 against rotation and under normal mode attack, which is very less as compared to rest of attacks used in this experiment.

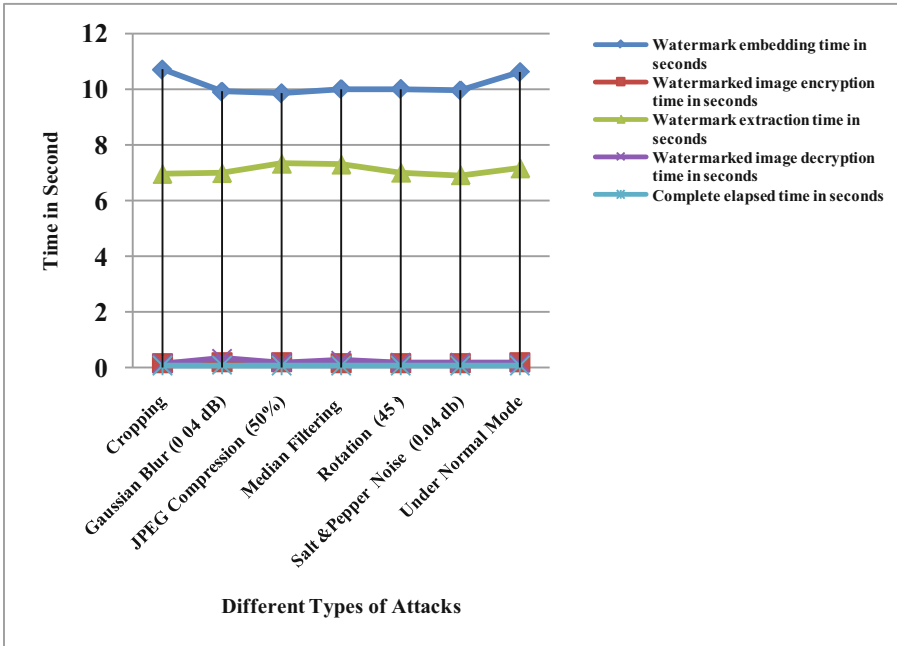


Fig. 9. Graphical presentation of time versus different attacks.

The resultant fuzzy watermarked cheque image using dynamic fuzzy inference system is shown in Fig. 10.

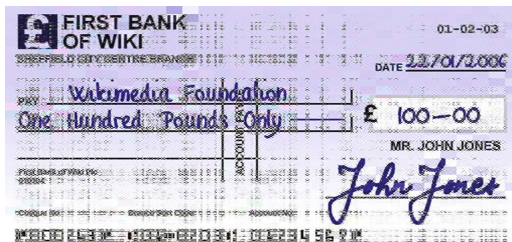


Fig. 10. Fuzzy watermarked cheque image using DFIS.

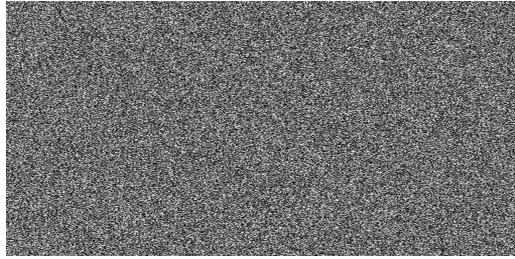


Fig. 11. Combined fuzzy watermarked and AES encrypted cheque image.

The different attacks are applied on Fig. 11, and it appears same against all attacks except rotation attack. After applying rotation attack with 45° on Fig. 11, the image appear as shown in Fig. 12.

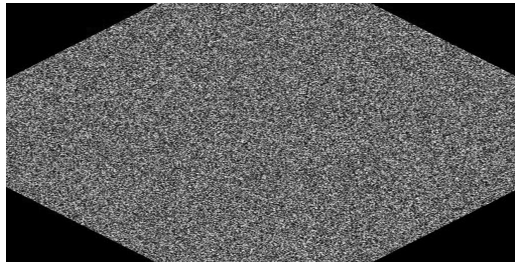


Fig. 12. Rotation attack combined fuzzy watermarked and AES encrypted cheque image.

The AES decryption process using 256 bits key is carried out on combined fuzzy watermarked and encrypted attacked image and the decrypted fuzzy watermarked cheque image obtained against various attack is as shown below in Figures.

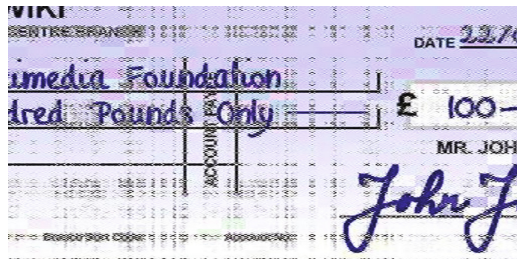


Fig. 13. Decrypted fuzzy watermarked cheque image obtained against cropping attack.

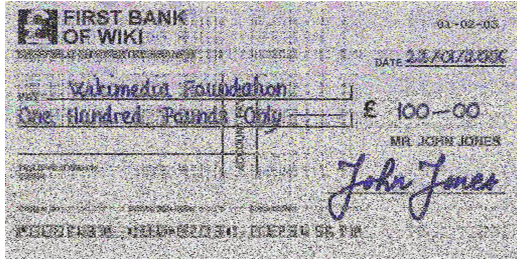


Fig. 14. Decrypted fuzzy watermarked cheque image obtained against Gaussian blur attack.

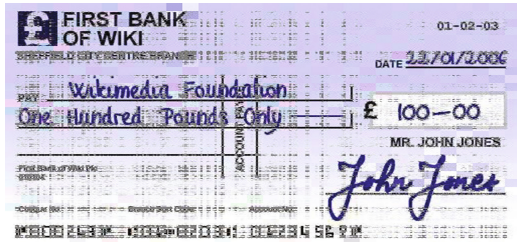


Fig. 15. Decrypted fuzzy watermarked cheque image obtained against JPEG Compression attack.

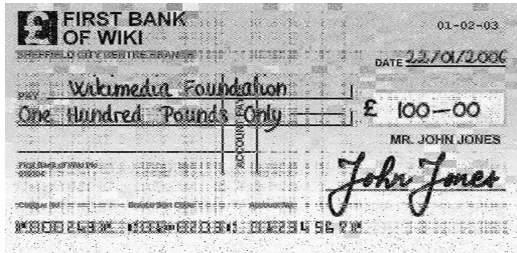


Fig. 16. Decrypted fuzzy watermarked cheque image obtained against Median filtering attack.

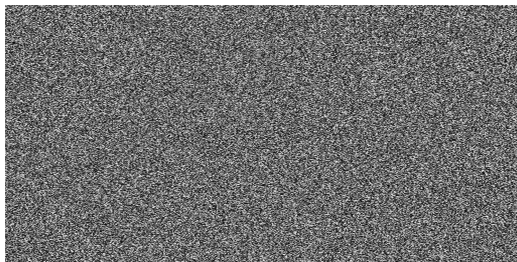


Fig. 17. Decrypted fuzzy watermarked cheque image obtained against Rotation attack.

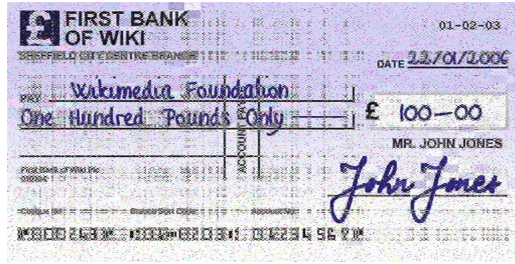


Fig. 18. Decrypted fuzzy watermarked cheque image obtained against Salt & Pepper noise attack.

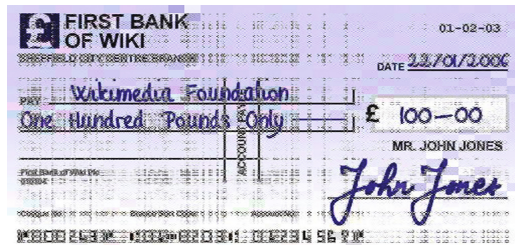


Fig. 19. Decrypted fuzzy watermarked cheque image obtained against under normal mode attack.

After achieving decrypted fuzzy watermarked cheque image, watermark process is carried out and watermark image is achieved. The extracted watermark image achieved against various attacks as shown below in Figs. 20 and 21.

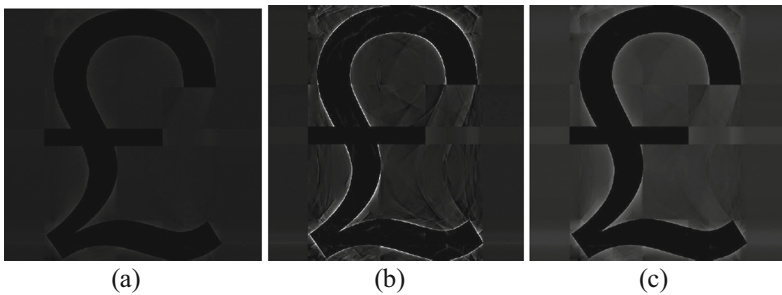


Fig. 20. Extracted watermark image obtained against (a) Cropping attack, (b) Gaussian blur attack, and (c) JPEG Compression attack.

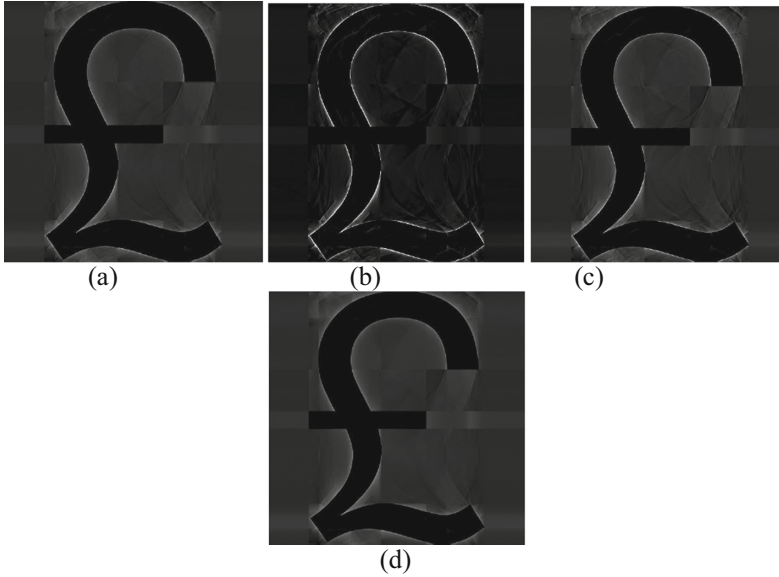


Fig. 21. Extracted watermark image obtained against (a) Median filtering attack, (b) Rotation attack (c) Salt & pepper noise attack, and (d) Under normal mode attack.

The Table 3 shows the values calculated for peak signal to noise ratio, mean square error for bank cheque image after watermarking and encryption as well as after decryption of cheque image and extraction of watermark, and normalized cross correlation coefficient i.e. robustness of watermark image against various attacks used in this experiment. From Table 3, it is observed that PSNR values of watermarked bank cheque image is 62.726 dB, which is constant against all attacks except cropping attack. The graphical representation shown in Fig. 22. It is also observed that the PSNR value of cheque image is decreased against all attacks. Similarly, the MSE value of watermarked cheque image is 0.0347 dB, which is constant against all attacks except cropping attack. However, the MSE values of bank cheque image gets increased after decryption & extraction of watermark as shown in Fig. 23. The Fig. 23 shows the graphical representation of mean square error and normalized cross correlation coefficient. It is observed that, the robustness of watermark after fuzzy watermarking and AES encryption is achieved 100% against attacks viz. (i) Cropping, (ii) JPEG Compression, and (iii) Under normal mode attack.

Table 3. PSNR value, MSE values and NCC values images against various attacks.

Different types of attacks	PSNR value of watermarked image in dB	PSNR value of image after extraction in dB	MSE value of watermarked image in dB	MSE of image after watermark extraction in dB	NCC value of watermark image	
					After encryption & watermarked image	After decryption & extraction of watermark logo
Cropping	57.704	49.903	0.1103	0.6640	1	0.8617
Gaussian Blur (0.04 dB)	62.726	50.280	0.0347	0.6090	0.9654	0.4857
JPEG Compression (50%)	62.726	51.013	0.0347	0.5140	1	0.8596
Median Filtering	62.726	50.963	0.0347	0.5200	0.999	0.8240
Rotation (45°)	62.726	49.878	0.0347	0.6680	0.0080	0.2340
Salt & Pepper Noise (0.04 db)	62.726	50.912	0.0347	0.5260	0.9330	0.7703
Under Normal Mode	62.726	51.013	0.0347	0.5140	1	0.8596

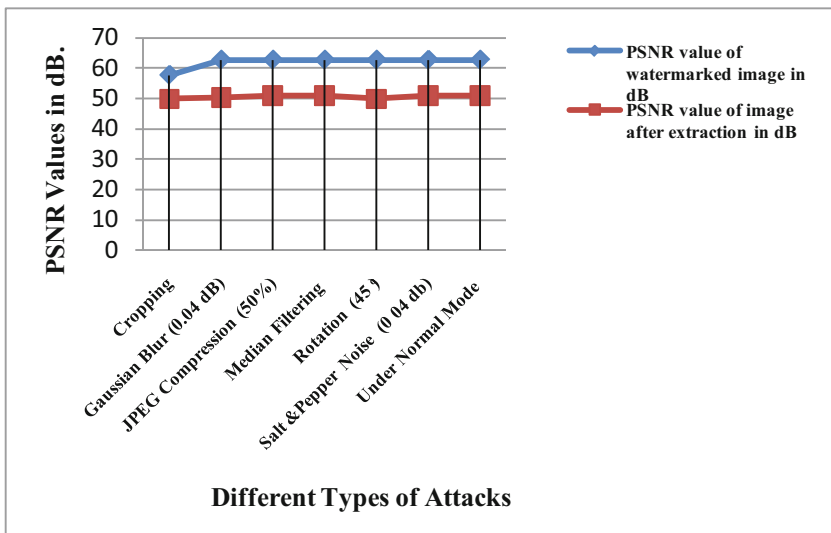


Fig. 22. Graphical presentation of PSNR values against different attacks.

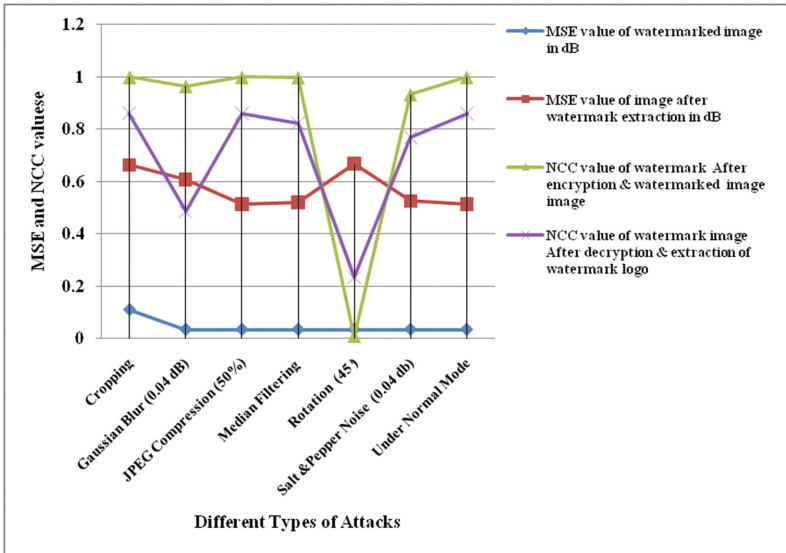


Fig. 23. Graphical presentation of NCC values of watermark and MSE values for cheque image against different attacks.

6 Conclusion

In this research paper, the combination of fuzzy logic and AES technique using 256 bits key is explained. The paper explains the DFIS used for watermarking of bank cheque and extraction of watermark for providing authentication and copyright protection service to digital bank cheque image. It also explains the AES technique using 256 bits key used for providing security services to watermarked bank cheque image. The robustness of watermark is achieved above 75% against maximum attacks except rotation attack with 45°. However, it fails to maintain robustness against rotational attack even after AES encryption using 256 bits key. The robustness of watermark achieved after extraction is maximum i.e. 86.17% against cropping attack. The robustness of watermark after extraction is found 85.96% against JPEG compression and under normal mode attack. It is observed that the robustness of extracted watermark is 23.40% obtained against rotation attack with 45°.

References

1. Lee, Z.Y., Yu, H.C., Kuo, P.J.: An analysis and comparison of different types of electronic payment systems. In: Proceedings of International Conference on Management of Engineering and Technology, (PICMET 2001), Portland, 29 July–2 August (2001)
2. Husain, F.: A survey of digital watermarking techniques for multimedia data. *Int. J. Electron. Commun. Eng.* **2**(1), 37–43 (2012)

3. Mahajan, P., Sachdeva, A.: A study of encryption algorithms AES, DES and RSA for security. *Glob. J. Comput. Sci. Technol. Netw.* **13**(15), 14–22 (2013). Online ISSN 0975-4172, Print ISSN 0975-4350
4. Metkar, S.P., Lichade, M.V.: Digital image improvement by integrating watermarking and encryption technique. In: Proceedings of the IEEE International Conference on Signal Processing, Computing and Control (ISPCC), 26–28 September (2013)
5. Gonge, S.S., Ghatol, A.A.: Combined DWT image watermarking and AES technique for digital 2-D Image. In: 2nd Proceedings of the IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2016), Jaipur, India, 23–25 December (2016)
6. Odeh, A., Masadeh, S.R., Azzazi, A.: A performance evaluation of common encryption techniques with secure watermark system (SWS). *Int. J. Netw. Secur. Appl.* **7**(3), (2015)
7. Reddy, V.R., Reddy, T.S.: Image encryption using fractional random wavelet transform. *Int. J. Adv. Res. Comput. Commun. Eng.* **3**(1), 4891–4893 (2014)
8. Ramamurthy, N., Varadarajan, S.: The robust digital image watermarking using quantization and fuzzy logic approach in DWT domain. *Int. J. Comput. Sci. Netw. ISSN 2277-5420*, **1**(5), (2012)
9. Ramamurthy, N., Varadarajan, S.: Robust digital image watermarking scheme with neural network and fuzzy logic approach. *Int. J. Emerg. Technol. Adv. Eng* **2**(9), (2012). ISSN 2250-2459
10. Lande, P.U., Talbar, S.N., Shinde, G.N.: Robust image adaptive watermarking using fuzzy logic an FPGA approach. *Int. J. Signal Process. Image Process. Pattern Recognit.* **3**(4), 43–54 (2010)
11. Ruanaidh, J.J.K.O., Pun, T.: Rotation, scale and translation invariant spread spectrum digital image watermarking. *Signal Process.* **66**(3), 303–317 (1998)
12. O'Ruanaidh, J.J.K., Pun, T.: Rotation, scale and translation invariant digital image watermarking. In: Proceedings of the International Conference on Image Processing, vol. 1. IEEE, (1997)
13. Gonge, S.S., Ghatol, A.: A cheque watermarking system using singular value decomposition for copyright protection of cheque images. In: Thampi, S.M., et al. (eds.) *Advances in Signal Processing and Intelligent Recognition Systems*. In: Proceedings of Advances in Intelligent Systems and Computing, vol. 425, Springer, Switzerland (2016)
14. Khalifa, O.O., Binti Yusof, Y., Abdalla, A.H., Olanrewaju, R.F.: State-of-the-art digital watermarking attacks. In: Proceedings of International Conference on Computer and Communication Engineering (ICCCE 2012), Kuala Lumpur, Malaysia, 3–5 July (2012)
15. Kundur, D., Hatzinakos, D.: A robust digital image watermarking method using wavelet-based fusion. In: Proceedings International Conference on Image Processing, vol. 1, IEEE (1997)
16. Tang, C.W., Hang, H.M.: A feature-based robust digital image watermarking scheme. *IEEE Trans. Signal Process.* **51**(4), 950–959 (2003)
17. Thapa, M., Sood, D.S.K., Sharma, A.P.M.: Digital image watermarking technique based on different attacks. *Int. J. Adv. Comput. Sci. Appl. (IJACSA)* **2**(4), 14–19 (2011)

Analysis of AES-GCM Cipher Suites in TLS

B. Arunkumar^(✉) and G. Kousalya

Coimbatore Institute of Technology, Coimbatore, India
aruncit17@gmail.com

Abstract. Encryption and decryption are the two most important complex methods for achieving security in any type of smart devices and systems/machines through transport layer security protocol (TLS). The symmetric key algorithms are the significant method for encrypting and decrypting the data/information using block cipher or stream cipher which is used for TLS protocol. The primary symmetric key block cipher algorithm used in TLS is Advanced Encryption standard (AES) and it provides security based on the key bits used in AES operation. The TLS protocol provides confidentiality(C), integrity (I) and Authenticity (A) in a single pass communication that is Authentication Encryption and Authentication Data (AEAD) between web browser and web server. It uses well known TLS cipher suite AES-GCM (Galois Counter mode) which is commonly used in TLS1.2. Suppose AES-NI hardware acceleration is not available in smart devices like tablets it causes performance issues in smart devices using TLS 1.2 protocol. If the smart device does not possess AES-NI, it can use software for running AES-GCM but it takes a lot of time for encryption/decryption of information, ergo causing the battery performance in smart devices. The newer symmetric Stream cipher CHACHA20-POLY1305 provides AEAD for securing the communication in smart devices thus reducing the battery cycles which is used for TLS 1.3. The paper discusses the pros and cons of AES-GCM authentication encryption used in TLS 1.2.

Keywords: AES-GCM · AEAD · TLS 1.2

1 Introduction

Transport layer security is the critical parameter in cyber world to secure the data between web browser and web server. The TLS protocol operates on two levels i.e. the TLS handshake and the TLS record [1]. The secret key and master secret key can be exchanged between web browser and web server using TLS handshaking method. The important concept in TLS is the Record layer where the exchanging of bulk data using symmetric key encryption between web servers to web browser. This is done so via secret key shared by the TLS handshake method. Encryption of data involves either, the use of symmetric block ciphers or the symmetric stream ciphers depending upon the hardware and software performance of the users systems or smart devices. In earlier history various TLS protocol levels used both block cipher and stream cipher based on the server and browser performance. In TLS protocol levels TLS 1.0, TLS 1.1, TLS 1.2 and TLS 1.3(Draft) gives better security and compatibility features for securing the data between web browser and web server. In symmetric-key block cipher encryption

methods, AES is the important cryptographic technique to be used in latest TLS 1.2 [2] protocol to secure the data. Once the data is encrypted and the subsequent essential factor for achieving the data Authentication between web browser and web server using MD5 [3], HMAC-SHA1 and HMAC-SHA 256,384,512 [4]. When providing encryption and authentication in two separate processes, achieving the security in browsers and servers will take more processing time. A new concept Authenticated encryption (AE) [5] method provides better security and processing time compare to the older methods was introduced. In the latest TLS 1.2 and 1.3 protocol levels using AE method to improve the security features compare to the older protocols TLS 1.0 and TLS 1.1. Emphasis on security features and processing time let the researchers to implement AES-GCM (Galois Counter mode) authenticated encryption method in TLS 1.2 protocol level. AES-GCM [6] is the technique used most of the modern browsers and servers to achieving the better security when compared to older block cipher and stream cipher cryptographic techniques. In AES-GCM provides both encryption and authentication as a parallizable method. The AES-GCM supports Authentication Encryption and Authentication Data (AEAD) [5] method in most of the cipher suites present in TLS 1.2 protocol. The Earlier TLS 1.0, TLS 1.1 cipher suites does not support Authentication encryption and AEAD. The earlier cipher suites providing the encryption using DES-CBC, AES-CBC and RC4-CBC and message authentication using HMAC-MD5, HMAC-SHA1, HMAC-SHA256 for achieving the security but it does so as two separate independent processes. Once AE introduced by researchers a number of authentication encryption techniques proposed like IAPM [25], XECB, OCB, CCM, EAX, CWC and GCM [7]. In AE techniques CWC [8] and GCM wins the competition of AE methods based on the following parameters provable security, parallelizability, high performance in hardware and software and unpatented. But researchers applying GCM in TLS 1.2 compare to CWC based on hardware performance that is GCM computes universal hash over GF (2^{128}) where CWC uses the prime field GF ($2^{127}-1$) that is complicated the hardware performance. So the AES-GCM is the important method used in most of the cipher suites used in TLS 1.2.

2 AES-GCM Description

AES [9] is an efficient block cipher encryption technique using most of the TLS cipher suites based on the criteria of security, cost, algorithm-implementation characteristics and Hardware and software efficiency. AES allows three different key lengths $\{0, 1\}^n$ where n varies among 128,192,256 bits used to encrypt the Plain text P belongs to $\{0, 1\}^n$ and produced the cipher text contains $\{0, 1\}^n$. The security of AES algorithm depends Number of rounds (Nr) used for each level of key length used in encrypting the data where Nr = 10 with key length of 128 bits, Nr = 12 with key length of 192 bits, Nr = 14 with key length of 256 bits.

2.1 GCM Description

Galois Counter mode (GCM) is mainly designed to attain paralyzing in authentication encryption technique. GCM performs two separate operations namely Encryption and

Authentication using a Block cipher AES. The Authentication operation calculates GHASH which uses Wegman-Carter polynomial (WMC) [10] Hashed over GF (2^{128}) illustrates in Fig. 1 and the encryption operation calculates GCTR using 128 bit AES counter mode operation illustrates in Fig. 2. Based on GHASH and GCTR, GCM produces the cipher text and Authentication Tag in a single-pass Authentication encryption for each session. The GCM [13] authentication encryption illustrates in Fig. 3 has four inputs and it is in the format of bit string given below,

- (1) A secret key K, key length depends on the AES
- (2) An Initialization vector IV, bits between 1 and 2^{64}
- (3) A plaintext P, bits between 0 and $2^{39}-256$
- (4) Additional Authentication Data (A), bits between 0 and 2^{64} and the two outputs are,
 - (i) Cipher Text C, length equivalent to plaintext length
 - (ii) Authentication Tag (T), bits between 0 and 128

2.2 GCM Notations

The block cipher encryption technique using GHASH function of the value X with the key K is denoted as E(K,X). The multiplication of two elements A, B GF (2^{128}) is denoted as A.B and the addition of A and B is A + B. The function len() returns 64-bit string containing positive integer describing the number of bits in its argument with the least significant bits occurred on the right side. The expression 0^L denotes a string of L zero bits and A||B denotes the concatenation of two bit strings A and B. The function MSB_t(S) returns the bit string containing only the most significant t bits of S.

The GHASH Function:

GHASH(X)

Steps:

1. Let $X_1, X_2, \dots, X_{M-1}, X_M$ denotes 128 bit sequence of blocks such that $X = X_1 || X_2 || \dots || X_{M-1} || X_M$
2. Let take $Y_0 = 0^{128}$
3. For $i = 1 \dots M$, Let $Y_i = (Y_{i-1} + X_i) \cdot H$
4. Return Y_M .

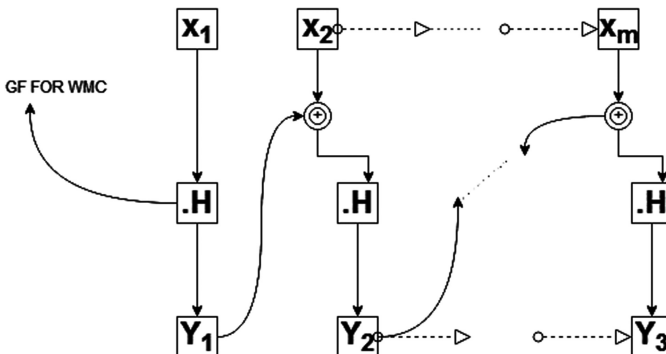


Fig. 1. GHASH Operation

The $\text{GHASH}_H(X)$ function can be expressed as

$$(X_1.H_M) + (X_2.H_{M-1}) + \dots\dots\dots(X_{M-1}.H_2) + (X_M.H)$$

The above condition optimizes the implementations of GHASH in both hardware and software machines.

The GCTR function:

GCTR (ICB, X)

Steps:

1. Let $n = \lceil \text{len}(X) / 128 \rceil$
2. Let $X_1, X_2, \dots, X_{n-1}, X_n^*$ denotes the unique sequence of 128 bit strings
Where $X = X_1 || X_2 \dots || X_{n-1} || X_n^*$ and X_1, X_2, \dots, X_{n-1} are complete blocks
3. Let $CB_1 = \text{ICB}$
4. For $i = 2$ to n , Let $CB_i = \text{inc}(CB_{i-1})$
5. For $i = 1$ to $n-1$, let $Y_i = X_i + \text{AES}_k(CB_i)$
6. Let $Y_n^* = X_n^* + \text{MSB}_{\text{len}(X_n^*)}(\text{AES}_k(CB_n))$
7. Let $Y = Y_1 || Y_2 || \dots || Y_n^*$
8. Return Y.

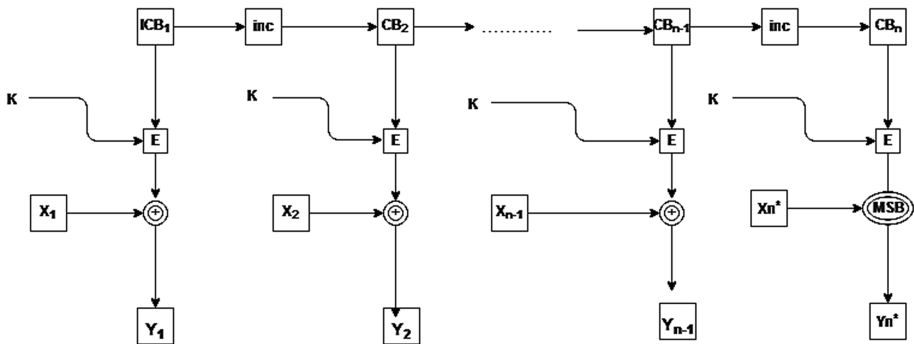


Fig. 2. GCTR Operation

The GCM-Authentication Encryption:

GCM-AES_k (IV, P, A)

Steps:

1. Let $H = \text{AES}_k(0128)$
2. Initialize a block J_0 , as follows
 - (i) If $\text{len}(IV) = 96$ then
 $J_0 = IV || 0^{31} 1$
 - (ii) If $\text{len}(IV) = 128$ then
 $J_0 = \text{GHASH}_H(IV || 0S)$
Where $S = 128 \cdot \lceil \text{len}(IV) / 128 \rceil - \text{len}(IV)$

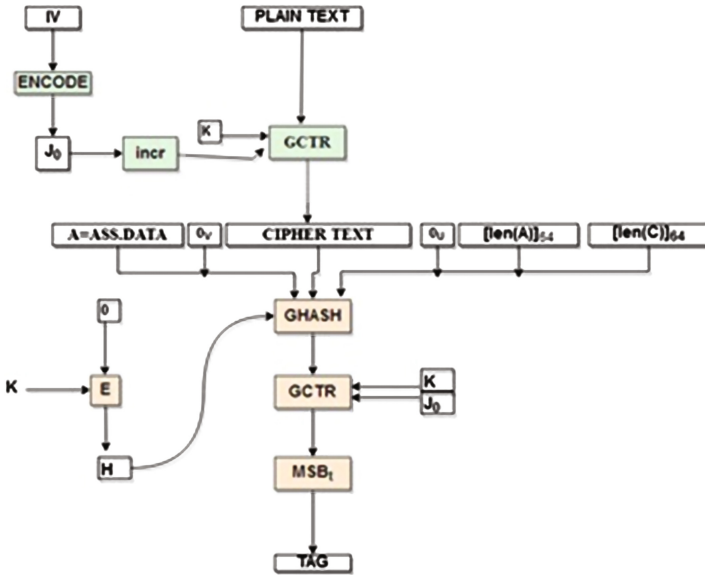


Fig. 3. GCM - authentication encryption

3. Let $C = GCTR_k(\text{inc}(J_0), P)$
4. Let $u = 128 \cdot \lceil \text{len}(C) / 128 \rceil - \text{len}(C)$ and $v = 128 \cdot \lceil \text{len}(A) / 128 \rceil - \text{len}(A)$
5. Define a block S as follows

$$S = \text{GHASH}_H(A || 0^v || C || 0^u || [\text{len}(A)]_{64} || [\text{len}(C)]_{64})$$
6. Let $T = \text{MSB}_t(GCTR_k(J_0, S))$
7. Return (C, T)

3 AES-GCM Performance in TLS

In TLS 1.2 cipher suites is the improved version of TLS 1.1 and eliminates the security problems and minimizes the processing time in TLS 1.1. based on this improvement most of the browsers and servers moved towards TLS1.2. TLS 1.2 Cipher suites introduced a new concept called AEAD with Authentication encryption using latest Cryptographic symmetric block cipher encryption technique. But number of authentication encryption methods (IAPM, CCM, CWC, GCM) [21] introduced over years achieving security using well known block cipher encryption AES [24]. Among these evolved AES technique, GCM gained priority for its efficiency in security.

3.1 Security of AES-GCM

The security of AES depends on the pseudorandom permutation and Nonce properly used in AES-GCM Authentication encryption. The security of GCM also depends on the block cipher AES. In GCM mode operation provides authenticity using GHASH

algorithm with the use of Wegman-Carter polynomial (WMC) Macs. So the Security bounds of the GHASH algorithm with n-bit tag will give $2^{-n/2}$ security against forgery [11] and also the IV value used in GHASH remains fresh for every session to provide better security in AES-GCM. The security of GCM analysis by two aspects is that privacy and authenticity. According to the privacy and authenticity in GCM, the adversary (A) should be infeasible to derive the cipher text (C) information without the access of secret key of Block cipher E. The adversary for authenticity should be infeasible to forge valid cipher text information without providing access to the secret key. Based on the GCM security analysis the privacy and authenticity is the important factors and it depends on the secret key used in the Block cipher (E). So the Block cipher E can be derived by a secure Pseudorandom Permutation (PRP). Based on the PRP, the block cipher E cannot be distinguished from a random permutation by an adversary and also cannot distinguished by choosing its inputs and view its outputs. So the adversary can choose either secret key or random permutation for breaking the security of GCM. For this scenario, the block cipher AES takes input $\{0,1\}^{128}$ and returns the output cipher text $\{0,1\}^{128}$ based on the input and output condition that the adversary derives the block cipher E. the block cipher E based on the randomly chosen secret key (SE) or random permutation function (SEC), for the cases the probability of the adversary to find the cipher text is 0.5 and K is the event that the adversary guesses the secret key SE and KC is the permutation function SEC [12]. So the Adversary true positive probability less than false probability.

$$A_E = P[K|S_E] - P[K^C|S_E^C] \quad (1)$$

So the probability of Adversary, AE is between 0 and 1.

The GCM authentication encryption oracle accepts input bit strings are N, A and M and returns the bit strings C and T. Similarly the decryption oracles accept the input N, A, C and T and return the P or special symbol FAIL. For maintaining the confidentiality and authentication in GCM, we use IND-CPA and IND-CCA. Based on the two security assumptions in GCM Encryption is secure if an adversary presented with these oracles cannot tell if they contain GCM with a randomly selected key or if C and T are derived a random function of the other inputs and the probability Adversary succeeds the above condition is 0.5. The privacy adversary A has access to the GCM encryption oracle or random-bits oracle based on the pseudo random function(PRF). So the GCM encryption mode oracle writes as ENC_k , input (N, A and P) and returns and random bits oracle ϵ input (N, A, M) and return. The privacy defined as

$$Adv_{GCM[E,\tau]}^{priv}(\underline{AAdef}) \Pr \left[K \xleftarrow{\$} \kappa : A^{Enc_K(\dots)} \Rightarrow 1 \right] - \Pr[A^{\$(\dots)} \Rightarrow 1] \quad (2)$$

The Authenticity Adversary A has encryption and decryption oracles defined as Enc_k and Dec_k ,

$$Adv_{GCM[E,\tau]}^{auth}(\underline{Adef}) \Pr \left[K \xleftarrow{\$} \kappa : A^{Enc_K(\dots), Dec_K(\dots)} \text{ forges} \right] \quad (3)$$

Based on the privacy and authenticity notations in GCM proves the provable security in GCM described below,

The privacy advantage of GCM at most

$$\frac{0.5(\sigma + q + 1)^2}{2^{128}} + \frac{2^{22}q(\sigma + q)(l_N + 1)}{2^{128}} \quad (4)$$

$$\frac{0.5(\sigma + q + q^1 + 1)^2}{2^{128}} + \frac{2^{22}(q + q^1 + 1)(\sigma + q)(l_N + 1)}{2^{128}} \quad (5)$$

For an Adversary (A) having q encryption and q^1 decryption oracles, where q is the total length of plain text which is of plaintext's at most blocks and q^1 represents the maximum nonce length of cipher text's at most blocks. The constant used in GCM authenticity and encryption value should not be less than 2^{20} .

3.2 AES-GCM Security Advantages

The AES-GCM security depends on the key values ($N = 128, 192$ and 256) used in AES algorithm and the security of AES-N-GCM [12] defined as,

If there are no attacks against AES-N that can distinguish from a random permutation with advantage greater than $AAES-N$, and no more than q packets are processed then,

$$A_{PRF} \geq A_{AES-N} + q^2 2^{-116} - q^2 2^{-89.4} \quad (6)$$

$$A_{PRP} \geq A_{AES-N} + q^2 2^{-116} - q^2 2^{-89.4} - q^2 2^{-128} \quad (7)$$

Based on the two conditions AES is indistinguishable from a random permutation function. The AES-GCM proves the security either by using PRF with randomly selected secret key (or) truly random function. The security analysis proved by GCM [6] is secure whenever the block cipher is indistinguishable from random values and condition described as,

$$A_{PRP} \geq q^2 l^2 2^{-142} + q^2 l^3 2^{-147} \quad (8)$$

4 Security Issues in AES-GCM

Although AES-GCM ensures maximum security, there are few constraints that degrade the efficiency of security. When the GHASH function used in GCM Authentication encryption [20], is computed by initializing $E_k(0) = H = 0^{128}$, the security of GHASH breaks down based on the powers of H (hash key) that repeats at short cycle. GHASH collisions can be achieved by adversary and produce the message forgery in GHASH Authentication function. For achieving the message forgery in GHASH functions, the WMC GF (2^{128}) having $2^9 = 512$ different multiplicative subgroups are involved in the computation. The GHASH operation is defined based on the finite field of GF (2^{128}) is,

$$Y_M = \sum_{i=1}^m X_i \times H^{m-i+1} \quad (9)$$

Based on Eq. (9), the adversary chooses GHASH collision operation and swap the two cipher text blocks named X_i and X_j achieving message forgery in GHASH operation [11].

Theorem: Let n be a number satisfying $\gcd(2^{128} - 1, n) = n$. Blindly swapping blocks X_i and X_j when $i \equiv j \pmod{n}$ will result successful forgery with probability of at least $n + 1/2^{128}$ for some random H .

Based on the theorem we have worked multiplicative groups for smoothing weak keys for binary finite field in GHASH operation. But this condition will not work in prime field $GF(P)$ with special Sophie German prime with the condition of $P = 2^{128} + 12451$ [14]. The next security issue in GCM authentication encryption is the adversary can compromise the secret key of the keyed hash function for achieving the security problem in GCM authentication using Chosen IV value attack. As GCM based counter mode operation uses standard 96 bit IV (Initial Vector) value will cause serious security problem in TLS 1.2 cipher suites. So to overcome this security issue, different length IV values can be added to uplift better security in GCM authentication. Another security issue in GCM is forbidden attack when using repeated IV values hence the adversary creates the authentication key without the knowledge of the master secret key and forges the cipher text. So the forbidden attack may be possible when choosing nonce value in GCM authentication, 96-bit nonce value, duplicate nonce and Random nonce. The next security issues in GCM authentication based on Ferguson's [22, 23] comments describes two weakness based on short authentication tag used in GCM authentication, the first one is, the probability of a successful forgery possible and the second weakness points to if the adversary reveals the authentication key if the adversary successfully forge the cipher text messages.

5 Hardware and Software Performance AES-GCM

The hardware and software performance of AES-GCM depends on the memory cycles and processing time needed for encrypting and decrypting the information in latest TLS 1.2 cipher suites. So the hardware performance of the AES algorithm depends on what type of processor used in machines or smart devices. But after the year 2010, most of the Intel processor supports new AES instruction set that is AES-NI and speeds up the memory cycles and processing time in TLS 1.2 cipher suites and the performance of the AES-NI compared to software is 3 to 10x better [15]. The following new instructions used in new Intel AES-NI [26, 27] processor is AESNC, AESENCLAST, AECDEC, AESDECLAST, AESKEYGENASSISI and AESIMC [16] and it improves the hardware performance and also provides the better security against the side channel attacks in AES. The hardware performance of the GCM authentication achieved by the new instruction developed by Intel that is PCLMULQDQ [16, 28] instructions uses binary polynomial multiplication and speeds up the computation in binary fields. When comparing the hardware performance of new AES-NI gives better results in memory

and processing time than pre AES-NI instruction set architecture. The performance of AES-GCM authentication encryption in newer AES-NI instruction set architecture achieves the pipeline with paralyzing control and the two functions GCTR and GHASH are interleaving by one function to improve the hardware performance in AES-NI. But older machines does not have new Intel AES-NI instruction set architecture and also smart devices implementing AES-GCM authentication encryption using software will degrade the performance of TLS 1.2 communication. Thus the researcher introduced new stream cipher based authentication encryption (CHACHA20-POLY1305) [17] cipher suite in TLS 1.3 [18, 19] which in turn improves the hardware and software performance in smart devices.

6 Conclusion

Since the performance and security features of AES-GCM authentication encryption spikes maximum, it is used in most of the TLS 1.2 cipher suites that is adopted in most modern web browsers and servers. The AES-GCM works better with the use of new Intel AES-NI and PCLMULQDQ instruction set architecture and resist the cache based side channel attacks when using new Intel set architecture. But most of the smart devices do not have Intel based architecture instruction set and hence it causes performance degradation and security issues when using TLS 1.2 Cipher suites. So this paper also suggests new stream cipher based authentication encryption technique CHACHA20-POLY1305 which is used in new TLS protocol TLS 1.3 will improve the hardware and software performance along with improvisation of security features in both systems and smart devices.

References

1. Bellare, M., Tackmann, B.: The multi-user security of authenticated encryption: AES-GCM in TLS1.3. In: *Advances in Cryptology—CRYPTO 2016*, pp. 247–276 (2016)
2. Meyer, C., Somorovsky, J., Weiss, E., Schwenk, J., Schinzel, S., Tews, E.: Revisiting SSL/TLS implementations: new bleichenbacher side channels and attacks. In: *23rd USENIX Security Symposium (USENIX 2014)*, pp. 733–748 (2014)
3. Krawczyk, H., Paterson, K.G., Wee, H.: On the security of the TLS protocol: a systematic analysis. In: *Advances in Cryptology—CRYPTO 2013*, pp. 429–448 (2013)
4. Wang, X., Yu, H.: How to Break MD5 and Other Hash Functions. In: Cramer, R. (ed.) *EUROCRYPT 2005*. LNCS, vol. 3494, pp. 19–35. Springer, Heidelberg (2005). doi:10.1007/11426639_2
5. Federal Information Processing Standards Publication 180-2. <http://csrc.nist.gov/publications/fips/fips180-2/fips180-2.pdf>
6. Rogaway, P., Atluri, V.: Authenticated-encryption with associated-data. In: *ACM Conference on Computer and Communications Security*, pp. 98–107 (2002)
7. McGrew, D.A., Viega, J.: The Galois/counter mode of operation (GCM). Submission to NIST modes of operation process. <http://csrc.nist.gov/CryptoToolkit/modes/proposedmodes> (2004)
8. Dworkin, M.: Recommendation for Block Cipher Modes of Operation: The CCM Mode for Authentication and Confidentiality. National Institute of Standards and Technology, NIST Special Publication 800-38C (2004)

9. Kohno, T., Viega, J., Whiting, D.: CWC: a high-performance conventional authenticated encryption mode. <http://eprint.iacr.org/2003/106/>
10. FIPS Pub. 197. Specification for the Advanced Encryption Standard (AES). National Institute of Standards and Technology, Federal Information Processing Standards (2001)
11. Bernstein, Daniel J.: Stronger Security Bounds for Wegman-Carter-Shoup Authenticators. In: Cramer, Ronald (ed.) EUROCRYPT 2005. LNCS, vol. 3494, pp. 164–180. Springer, Heidelberg (2005). doi:[10.1007/11426639_10](https://doi.org/10.1007/11426639_10)
12. Saarinen, M.O.: Cycling attacks on GCM, GHASH and other polynomial MACs and hashes. In: Canteaut, A. (ed.) FSE 2012. LNCS, vol. 7549, pp. 216–225. Springer, Berlin (2012)
13. McGrew, D.A., Viega, J.: The security and performance of the galois/counter mode of operation (full version). Cryptology ePrint Archive, Report 2004/193 (2004). <http://eprint.iacr.org/>
14. Iwata, T., Ohashi, K., Minematsu, K.: Breaking and repairing GCM security proofs. Cryptology ePrint Archive, Report 2012/438 (2012). <http://eprint.iacr.org/>
15. Saarinen, M.O.: SGCM: the Sophie Germain counter mode. Cryptology ePrint Archive, Report 2011/326 (2011). <http://eprint.iacr.org/>
16. Gueron, S., Kounavis, M.E.: Intel Carry-Less Multiplication Instruction and its Usage for Computing the GCM Mode (Rev. 2). Intel Software Network (2010). <http://software.intel.com/en-us/articles/carry-less-multiplication-and-its-usage-for-computing-the-gcm-mode/>
17. Gueron, S., Krasnov, V.: [PATCH] efficient implementation of AES-GCM, using Intel's AES-NI, PCLMULQDQ instruction, and the advanced vector extension (AVX). <http://rt.openssl.org/Ticket/Display.html?id=2900>. Accessed Oct 2012
18. Procter, Gordon: A Security analysis of the composition of ChaCha20 and Poly1305. IACR Cryptol. ePrint Arch. **2014**, 613 (2014)
19. A cryptographic analysis of the TLS 1.3 draft-10 full and pre-shared key handshake protocol (2016). <http://eprint.iacr.org/2016/081>
20. Yap, W., Yeo, S.L., Heng, S., Henricksen, M.: Security analysis of GCM for communication. Secur. Commun. Netw. **7**(5), 854–864 (2014)
21. Andreeva, E., Bogdanov, A., Luykx, A., Mennink, B., Tischhauser, E., Yasuda, K.: Parallelizable and Authenticated Online Ciphers. In: Sako, K., Sarkar, P. (eds.) ASIACRYPT 2013. LNCS, vol. 8269, pp. 424–443. Springer, Heidelberg (2013). doi:[10.1007/978-3-642-42033-7_22](https://doi.org/10.1007/978-3-642-42033-7_22)
22. Bellare, M., Rogaway, P., Wagner, D.: The EAX Mode of Operation. In: Roy, B., Meier, W. (eds.) FSE 2004. LNCS, vol. 3017, pp. 389–407. Springer, Heidelberg (2004). doi:[10.1007/978-3-540-25937-4_25](https://doi.org/10.1007/978-3-540-25937-4_25)
23. Ferguson, N.: Authentication weaknesses in GCM. NIST Comment (2005)
24. Bernstein, D.J.: Cache-timing attacks on AES. Technical report, 2005 Antoine Joux. Authentication failures in NIST version of GCM (2006). http://csrc.nist.gov/groups/ST/toolkit/BCM/documents/Joux_comments.pdf. Accessed 20 Feb 2016
25. Hastad, J.: The security of the IAPM and IACBC modes. J. Cryptol. **20**(2), 153–163 (2007)
26. Akdemir, K. e.a.: Breakthrough AES performance with intel AES new instructions, Intel Whitepaper (2010). <http://software.intel.com/file/27067>
27. Gopal, V. et al.: Optimized Galois-counter-mode implementation on intel architecture processors, Intel Whitepaper (2010). <http://download.intel.com/design/intarch/PAPERS/324194.pdf>
28. Hoban, A.: Using intel AES new instructions and PCLMULQDQ to significantly improve IPsec performance on Linux, Intel Whitepaper (2010), <https://www.Intel.com/design/intarch/papers/324238.pdf>

Data Classification Using Machine Learning Approach

Shekhar Pandey^(✉), Supriya M, and Abhilash Shrivastava

Department of Computer Science and Engineering,
Amrita School of Engineering, Amrita Vishwa Vidyapeetham,
Amrita University, Bengaluru, India
chirag1989shekhar@gmail.com,
m_supriya@blr.amrita.edu,
shrivastava.abhilash2510@gmail.com

Abstract. Currently, Internet has numerous effects on our everyday lifecycle. Its significance as an intermediate for commercial transactions will develop exponentially throughout the next years. In terms of the engaged marketplace volume, the Business to Business region will hereby be the supreme exciting area. As the extensive usage of electronic business transactions increase, great volume of products information gets generated and managing such large information automatically becomes a challenging task. The accurate classification of such products to each of the existing classes also becomes an additional multifarious task. The catalog classification is an essential part for operative electronic business applications and classical machine learning problems. This paper presents a supervised Multinomial Naïve Bayes Classifier machine learning algorithm to classify product listings to anonymous marketplaces. If the existing products are classified under the master taxonomy, the task is to automatically categorize a new product into one of the existing categories. Our algorithm approach proposes a method to accurately classify the existing millions of products

Keywords: Naïve Bayes · Classifier · Machine learning · Categories

1 Introduction

Small scale to giant scale sized businesses who trade products online spend a substantial part of their time, money, and struggle - in categorizing the products they trade, to better market their products, and in determining which products to sell. Such E-inventory (Electronic index) businesses hold their data of items and administrations in a web based business association. E-list is a type of classification in which information of an inventory is categorized to one of the already existing classes list. The classes are categorized by a definite taxonomy framework which as a rule has an arranged structure. Accurate taxonomy is essential not just for information introduction and synchronization among business accomplices, additionally to keep the quickly expanding item information viable and adequate. Still, merchandise data cataloguing is an extremely tedious job and not easy to do by hand due to its enlarged product information. In this

paper, the use of automatic learning techniques has been proposed to outline product classes (e.g., ‘Electronics’) and potential subcategories (e.g., ‘Printers’). This is beneficial for the circumstance where a business has a list of new products that has to be sold by automatically classifying based on training data of the businesses’, other products and classifications. This will also be beneficial for categorizing a new merchandise item line that has not been previously introduced in the market before, or for the items that are more densely populated than the training data set. To advance this procedure, an automatic learning algorithm has been proposed that can automatically categorize listings with high accuracy. A number of competitor customary classification schemes are already available in the market but none of them are globally recognized and accepted [1]. Works in [2] connected a few procedures from data recovery and machine learning approaches, figuring out how to proceed with item information characterization by incorporating striking calculations like KNN (K-Nearest Neighbor), SVM (Support Vector Machine) and NBC (Naïve Bayes classifier) etc.

2 Related Works

Currently, with the fast development of small-organized documents, their taxonomy categorization issue has pulled in an expanding consideration. The normal motivation is that the arrangement of such large documents may comprehend helpful data for classification. There is a need for several attempts to analyze anonymous marketplaces [3]. In [3], the author analyses on the Silk Road for 8 months, investigating product entries and the complete distribution of product listings. Be that as it may, they depended on seller provided classifications, which does not exist for all commercial centers. Also, a few sellers purposefully misclassify their item to seem higher in commercial center query output. Correcting for these problems, the categorization of taxonomy [4] has been constructed on the Bayesian networks. For each document in the preparation set, it amasses a Bayesian system whose structure is basically the same as that of the record itself appeared as a tree. Constructed on Bayesian networks conditional probability, the work proposed in [4] develops the classification of an information archive. Zaki et al. [5] describe a technique for construction for XML taxonomy classifier based on administrator and Denoyer et al. [6] recommend a classifier which classifies structured multimedia taxonomy based on Bayesian. Vinithra et al. [7], Ani et al. [8] and Priyanka et al. [9] also outlines the various classification techniques. Most of the automatic learning methods are well documented in the literature as effective binding blocks for document classification systems.

Motivated by the study from different researchers, this work decides to deal with Multinomial Naïve Bayes Classifier to making a document classifier.

Naïve Bayes approach works on each word position which is described to be an attribute of the Naïve Bayes Classifier [10, 11] for level content classification. Moreover, seeing that attribute has distinctive frequency power individually, we indulgence singular attributes diversely by allotting weights as per their significance. Each attributes are normalized before allotting the weights to the attributes sensibly. The frequency method has been used for exactness classification. Our classifier demonstrates enhanced exactness with the Multinomial Naïve Bayes Classifier even if there is noisy data.

3 Classification Algorithm Multinomial Naïve Bayes Classifier

The Multinomial Naïve Bayes is adequately automatic calculation for content classification because of its quality and has good execution performance. We demonstrate this methodology to classify the catalogs rendering to their classes.

3.1 Multinomial Naïve Bayes Classifier

To categorize an item or word w , the Naïve Bayes calculates the posterior probability $P(w|d)$ of that word or item constructed on the Bayes Theorem. Specified a group of classes Z , the attributes $\langle k_1, k_2, \dots, k_n \rangle$ and the values $\langle f_1, f_2, \dots, f_n \rangle$ that designate an input instance, the Naïve Bayes allocates the most likely classification as specified by the supplementary calculation method.

$$Z_{NB} = \arg \max_{w_j \in Z} P(w_j) \prod_i P(k_i = f_i | w_j) \quad (1)$$

where Z_{NB} is Naïve Bayes class.

This method classifies any catalogs which have huge number of attributes based on the word's probability and frequency. This model does not neglect any words even if they have less probability because this model treats all words equally to get accurate results. When this method is used for catalog classification, every written text treats as attribute for classification. Specified each word m , treat as individually $\langle m_1, m_2, \dots, m_n \rangle$ that constitutes an input document, the Naïve Bayes can be represented as

$$Z_{NB} = \arg \max_{w_j \in Z} P(w_j) \prod_i P(k_i = m_i | w_j) \quad (2)$$

Treating that each word has equal priority and supposition that the elements are indistinguishably conveyed to reduce the cost, the above approach implies that the probability of experiencing each word is independent of the particular word position [4].

3.2 Extending and Normalizing Attributes

Since the qualities of writings are made out of many words and are regularly boisterous, tolerating just the correct matches is deluding. It is plainly wrong to recognize "Laptop" and 'Laptop Notebook'. The issue ends up being more lamentable when we endeavor to use a property like 'item portrayal' which is now and again made out of full sentences. So, sometimes even each word method is not successful because of the same name with different writing styles. Hence, we reclassify the estimation of a property as

$$f_i = \{n_{i1}, n_{i2}, \dots, n_{iq}\} \quad (3)$$

where n_{iq} is a value formed from f_i by the parser. Then we can reasonably assume that

$$\begin{aligned}
 Z_{NB} &= \arg \max_{w_j \in Z} P(w_j) \prod_{i,j} P(n_{iq} \text{ appears in } k_i | w_j) \\
 &= \arg \max_{z \in Z} \left\{ |w_j| \prod_{iq} \frac{n(w_j, k_i, n_{iq})}{n(w_j, k_i)} \right\}
 \end{aligned}
 \tag{4}$$

where $n(w_j, k_i, n_{iq})$ is the existences of n_{iq} in k_i of the phrase which indicates of class w_j . Similarly, $n(w_j, k_i)$ is the total frequencies of all in k_i of the catalogs that belong to class w_j . $n(k_i)$ is the total number of words in k_i (Figs. 1 and 2).

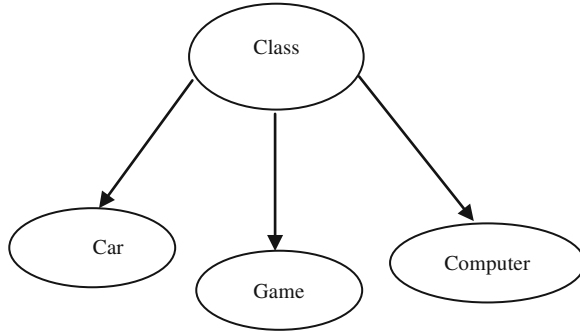


Fig. 1. This showing how we can classify class with unique name from a dataset

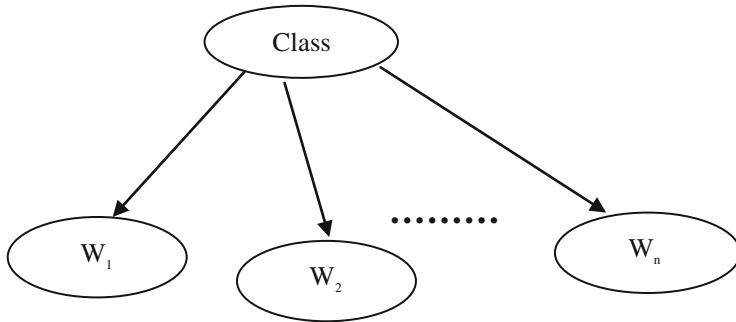


Fig. 2. This show how it will work by taking each word in calculation

The above model works perfectly when text phrases are small and not much bigger, but when some text phrases are big then they will generate more high frequency as compared to small phrases, so we need to overcome it by using the following equation.

$$\begin{aligned}
 Z_{NB} &= \arg \max_{w_j \in Z} P(w_j) \prod_i \left(\prod_q P(n_{iq} \text{ appears in } k_i | w_j) \right)^{\frac{1}{|k_i|}} \\
 &= \arg \max_{w \in Z} \left\{ |w_j| \prod_i \left(\prod_q \frac{n(w_j, k_i, n_{iq})}{n(w_j, k_i)} \right)^{\frac{1}{|k_i|}} \right\}
 \end{aligned}
 \tag{5}$$

We used the geometric mean for the normalization and after applying the mean the final equation is given as

$$\begin{aligned} Z_{NB} &= \arg \max_{w_j \in Z} P(w_j) \prod_i \left(\prod_q P(n_{iq} \text{ appears in } k_i | w_j) \right)^{\frac{m_i}{|V_i|}} \\ &= \arg \max_{w_j \in Z} \left\{ |w_j| \prod_i \left(\prod_q \frac{n(w_j, k_i, n_{iq})}{n(w_j, k_i)} \right)^{\frac{m_i}{|V_i|}} \right\} \end{aligned} \quad (6)$$

where m_i is the weight of the attribute of k_i .

3.3 Applying Multinomial Naïve Bayes Classifier

To change over the test to elements, we used the frequency changes over every token in the listing weight in order to find out how vital that token is to listing; standardized by the quantity of times the token shows up in the entire corpus. This normalization diminishes the effect of basic token in the corpus. To discover any item that has a place with its class, the following steps are to be implemented:

Step 1. Compute the prior probabilities

$$P(\text{category}) = \frac{\text{Number of records classified into the category}}{\text{Total number of the records}}$$

Step 2. Compute likelihood.

$$\begin{aligned} P(\text{word}/\text{category}) &= \frac{\text{Number of frequency of a word in all records from a category} + 1}{\text{All the words in every document from a category} + \text{total number of unique words in all the records}} \end{aligned}$$

Step 3. Final computation

$$\begin{aligned} P(\text{category}/\text{records}) &= \\ &P(\text{category}) * P(\text{word}_1/\text{category}) * P(\text{word}_2/\text{category}) * \dots * \\ &P(\text{word}_n/\text{category}) \end{aligned}$$

The product belongs to the class that has the highest probability among others.

4 Implementation Multinomial Naïve Bayes Classifier

4.1 Training (Step 1 and Step 2 from Sect. 3.2)

While training the dataset of different classes, we count each word as individual and then form the frequency and probability based on number of same occurrence of that word. The dataset Fig. 3 shows that there are more than two classes like Car, Game, and Computer.

Index	Index	Content	Class
0	0	Saturn Merchant Auto	Car
1	1	Toyota Auto Mercedes	Car
2	2	Football Game Play	Game
3	3	Cycle Excercie Game	Game
4	4	Laptop PC Mouse	Computer

Fig. 3. Sample dataset of products which has three classes

4.2 Dataset

For clear description and ease of presence of the method, a test dataset has been shown in Fig. 3. But, the size of the actual dataset is 36256 KB and has been used for validation of the method. The actual dataset can be found in the link (<https://github.com/sam-chirag/Data-Classification-Using-Machine-Learning-Dataset>)

After applying above Multinomial Naïve Bayes Classifier method (step 1 and step 2) as given in Sect. 3.2, we will get frequency table with their probability as shown in Fig. 5.

4.3 Classification

An example of simple phrase is given below based on above training dataset shown in Fig. 4.

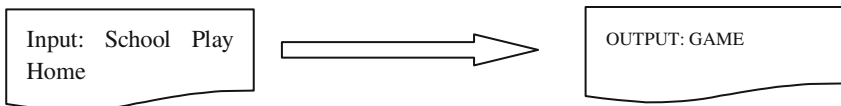


Fig. 4. Input catalog for classification

Classes: Car, Game, Computer

Class Car Probability:

$$P(Car/S1) = P(Car) * P(School/Car) * P(Play/Car) * P(Home/Car)$$

$$P(Car) = \frac{2}{5} = 0.4$$

$$P(School/Car) = \frac{(0 + 1)}{(6 + 13)} = 0.05263$$

$P(\text{Play}/\text{Car}) = 0.05263$ (This word exist in the wordlist of car and we directly taken its likelihood probability value from that table as shown in Fig. 5)

$$P(\text{Home}/\text{Car}) = \frac{(0 + 1)}{(6 + 13)} = 0.05263$$

So,

$$P(\text{Car}/S1) = 0.4 * 0.05263 * 0.05263 * 0.05263 = 0.00005825$$

Class Game Probability:

$$P(\text{Game}/S1) = P(\text{Game}) * P(\text{School}/\text{Game}) * P(\text{Home}/\text{Game}) * P(\text{Computer}/\text{Game})$$

$$P(\text{Game}) = \frac{2}{5} = 0.4$$

$$P(\text{School}/\text{Game}) = \frac{(0 + 1)}{(6 + 13)} = 0.05623$$

$$P(\text{Play}/\text{Game}) = 0.10$$

$$P(\text{Home}/\text{Game}) = \frac{(0 + 1)}{(6 + 13)} = 0.05623$$

So,

$$P(\text{Game}/S1) = 0.4 * 0.05623 * 0.105 * 0.05623 = 0.00011633$$

Class Computer Probability:

$$P(\text{Computer}/S1) = P(\text{Computer}) * P(\text{School}/\text{Computer}) * P(\text{Play}/\text{Computer}) * P(\text{Home}/\text{Computer})$$

$$P(\text{Computer}/S1) = \frac{1}{5} = 0.2$$

$$P(\text{School}/\text{Computer}) = \frac{(0 + 1)}{(3 + 13)} = 0.0625$$

$$P(\text{Play}/\text{Computer}) = 0.0625$$

$$P(\text{Home}/\text{Computer}) = \frac{(0 + 1)}{(3 + 13)} = 0.0625$$

So,

$$P(\text{Computer}/S1) = 0.2 * 0.0625 * 0.0625 * 0.0625 = 0.000048828$$

So as the probability of the given input among Game class is high, this text belongs to the Game class.

Similarly, the probability calculations for the Game Class and Computer Class has been performed and the results obtained are [0.0526, 0.105, 0.105, 0.105, 0.150, 0.0526, 0.0526, 0.0526, 0.0526, 0.0526, 0.105, 0.0526, 0.0526] and [0.0625, 0.0625, 0.0625, 0.0625, 0.125, 0.0625, 0.0625, 0.125, 0.125, 0.0625, 0.0625, 0.0625] respectively.

Figure 6 shows the final calculation result of the considered input set. The corresponding algorithms are given in Table 1.

Index	Word	Frequency	Count Words	Probability
0	Baseball	0	6	0.0526
1	Car	2	6	0.158
2	Colored	0	6	0.0526
3	Dealer	1	6	0.105
4	GIFs	0	6	0.0526
5	Game	0	6	0.0526
6	Muscle	0	6	0.0526
7	Play	0	6	0.0526
8	Pulled	0	6	0.0526
9	Root	0	6	0.0526
10	Saturn	1	6	0.105
11	Tercel	1	6	0.105
12	Toyota	1	6	0.105

Fig. 5. Car class with their frequency, count words (total words in car class), probability from the catalog in Fig. 3

5 Results and Conclusion

```
*****
searching concepts starts here
*****

Enter any string for class search

School Play Home

[School', 'Play', 'Home']
////////////////////////////////////
Car2.csv
////////////////////////////////////
[School', 'Play', 'Home']
['Saturn', 'Merchant', 'Auto', 'Toyota', 'Mercedes', 'Football', 'Game', 'Play', 'Cycle', 'Excercie', '',
'Laptop', 'PC', 'Mouse']
*****
School
*****
Play
*****
Home
{'School': -1, 'Play': 7, 'Home': -1}
6
0.05
6
0.05
////////////////////////////////////
Game2.csv
////////////////////////////////////
[School', 'Play', 'Home']
['Football', 'Game', 'Play', 'Cycle', 'Excercie', '', 'Saturn', 'Merchant', 'Auto', 'Toyota', 'Mercedes',
'Laptop', 'PC', 'Mouse']
*****
School
*****
Play
*****
Home
{'School': -1, 'Play': 2, 'Home': -1}
7
0.047619047619047616
7
0.047619047619047616
////////////////////////////////////
Computer2.csv
////////////////////////////////////
[School', 'Play', 'Home']
['Laptop', 'PC', 'Mouse', 'Saturn', 'Merchant', 'Auto', 'Toyota', 'Mercedes', 'Football', 'Game', 'Play',
'Cycle', 'Excercie', '']
*****
School
*****
Play
*****
Home
{'School': -1, 'Play': 10, 'Home': -1}
3
0.058823529411764705
3
0.058823529411764705
{'Car': 5.0000000000000016e-05, 'Game': 8.63837598531476e-05, 'Computer': 4.0708324852432325e-05}
*****
*****
Game
C:/Users/Samz (Sam)/Desktop/Testing Amazon dataset/newtest.py:391: DeprecationWarning: 'U' mode is deprecated
with open(class_name, 'rU') as infile:
```

INPUT

OUTPUT

Fig. 6. Calculation result of above input

Table 1. Classification algorithm

Training and Applying Multinomial Naïve Bayes Algorithm

(a) TRAINMULTINOMIALNAIYEBAYES(C, F)

- (1) $P \leftarrow \text{FINDVOCABULARY}(F)$
- (2) $N \leftarrow \text{COUNTDOCUMENTS}(F)$
- (3) for each $c \in C$
- (4) do $K \leftarrow \text{COUNTDOCUMENTSINCLASS}(F, c)$
- (5) $\text{prior}[c] \leftarrow K_c/K$
- (6) $\text{word}_c \leftarrow \text{CONCATTEXTOFALLDOCUMENTSINCLASS}(F, c)$
- (7) for each $j \in P$
- (8) do $J_{cj} \leftarrow \text{COUNTTOKENSOFTERM}(\text{word}_c, j)$
- (9) for each $j \in P$
- (10) do $\text{condprob}[j][c] \leftarrow \frac{J_{cj}+1}{\sum_j (J_{cj}+1)}$
- (11) return $P, \text{prior}, \text{condprob}$

(b) APPLYMULTINOMIALNAIYEBAYES(C, P, prior, condprob, d)

- (1) $T \leftarrow \text{EXTRACTWORDTOKENFROMDOCS}(P, d)$
- (2) for each $i \in C$
- (3) do $\text{score}[i] \leftarrow \log \text{prior}[i]$
- (4) for each $j \in T$
- (5) do $\text{score}[i] += \log \text{condprob}[j][i]$
- (6) return $\arg \max_{i \in C} \text{score}[i]$

Our experiments are carried out from product databases of Amazon, Flipkart, Snapdeal and Paytm. The database currently contains 40000 product catalogs and classification structure contains 1000 leaf classes. This experiment has been carried out on Intel Core i3 1.80 GHz machine which has 4 GB of RAM. Database server used is MySQL and the application software for implementation of programming code is Anaconda (Spyder 3.6). Approximately 70% accurate results were obtained based on our algorithm and it manages attribute-wise distribution of terms to adapt to the organized way of e-lists. The best thing is that with the help of normalization, our method without giving weightage to long text is able to give better results. We are in the process of improving the accuracy hence obtained. The algorithm could be made more powerful by including information from more sources. Test information drawn from a more extensive source would likewise give a superior speculation estimate.

References

1. Fensel, D., Ding, Y., Schulten, E., Omelayenko, B., Botquin, G., Brown, M., Flett, A.: Product data integration in B2B e-commerce. *IEEE Intell. Sys.* **16**(3), 54–59 (2001)
2. Ding, Y., Korotkiy, M., Omelayenko, B., Kartseva, V., Zykov, V., Klein, M., Schulten, E., Fensel, D.: GoldenBullet: automated classification of product data in e-commerce. In: *Business Information System* (2002)
3. Branwen, G.: Silk road: theory and practice (2015). <http://www.gwern.net/Silk%20Road>, Accessed 09 Dec 2015
4. Denoyer, L., Gallinari, P.: Bayesian network model for semi-structured document classification. *Inf. Process. Manag. (Elsevier)* **40**(5), 807–827 (2004)
5. Zaki, M.J., Aggarwal, C.C.: XRules: an effective structural classifier for XML data. In: 9th ACM SIGKDD, pp. 316–325 (2003)
6. Denoyer, L., Vittaut, J., Gallinari, P., Brunessaux, S., Brunessaux, S.: Structured multimedia document classification. In: *ACM DOCENG 2003*, pp. 153–160 (2003)
7. Vinithra, S.N., Anand Kumar, M., Soman, K.P.: Analysis of sentiment classification for Hindi movie reviews: a comparison of different classifiers. *Int. J. Appl. Eng. Res.* **10** (2015)
8. Ani, R., Sasi, G., Sankar, U.R., Deepa, O.S.: Decision support system for diagnosis and prediction of chronic renal failure using random subspace classification. In: *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 1287–1292. Jaipur (2016)
9. Priyanka, C., Gupta, D.: Fine grained sentiment classification of customer reviews using computational intelligent technique. *Int. J. Eng. Technol.* **7**(4), 1453–1468 (2015)
10. Sebastiani, F.: Machine learning in automated text categorization. *ACM Comput. Surv.* **34**(1), 1–47 (2002)
11. Mitchell, T.: *Machine Learning*. McGraw-Hill, Columbus (1997)

Topic Modeling for Unsupervised Concept Extraction and Document Ranking

V.S. Anoop¹(✉), S. Asharaf², and P. Deepak³

¹ Data Engineering Lab, Indian Institute of Information Technology and Management - Kerala, Thiruvananthapuram, India

anoop.res15@iiitmk.ac.in

² Indian Institute of Information Technology and Management - Kerala, Thiruvananthapuram, India

asharaf.s@iiitmk.ac.in

³ Queens University, Belfast, UK

d.padmanabhan@qub.ac.uk

Abstract. This paper proposes a framework which induces semantically rich concepts from probabilistically generated topics by a topic modeling algorithm. In this method an off-the-shelf tool has been used to extract noun-phrases as word bi-grams and tri-grams from the static document corpus and then models the topics using Latent Dirichlet Allocation algorithm. Additionally, we show that a small extension to our proposed framework can better rank documents in a large collection, which is a well studied area in information retrieval. Experiments conducted on three real world datasets show that this proposed framework outperforms state-of-the-art methods used for extracting concepts and ranking documents. When compared with the baselines chosen, our proposed concept extraction method showed an increased f-measure in the range of 16.65% to 22.04% and the proposed topic modeling guided document retrieval method showed 7.6%–16.61% increase in f-measure.

Keywords: Topic modeling · Concept extraction · Document ranking · Latent Dirichlet Allocation

1 Introduction

Topic modeling algorithms such as Latent Dirichlet Allocation (LDA) [1], Probabilistic Latent Semantic Indexing (PLSI) [2] and Probabilistic Latent Semantic Analysis (PLSA) [3] are proved to be extensively useful in bringing out hidden themes from textual content which are used for further analysis. Given a large collection of text documents, these algorithms generate such themes by representing each document as a random mixture over latent topics and assumes each topic be a probabilistic distribution over words. This idea is used in various scenarios such as document analysis and pattern identification in text mining,

because text is considered to be the best data on which topic modeling can be easily applied. Majority of the topic modeling algorithms work on the assumption called “bag-of-words” to generate “topics”. These statistically generated “topics” are sequences of word unigrams. The interpretation of such topics still remain as a hindrance to most of the text mining practitioners [4]. In real life scenarios, humans understand the concepts and key-phrases as a combination of words such as bi-grams and tri-grams. Splitting these bigrams and trigrams into unigrams cause its semantic meaning to be lost and that would generate ambiguous and irrelevant topics. For example, the concept “network security attack” is more interpretable for humans compared to the word unigram - “attacks”. Chopping them into “network”, “security” and “attack” unigrams cause its semantic meaning to be lost.

Concept mining and extraction plays a key role in text mining where each concept is a combination of words which are real or imaginary. Those concepts are used for tasks such as document retrieval [5,6], classification [7], concept based sentiment analysis [8], semantic product search in e-commerce [9,10] and concept hierarchy learning [21]. Search engines and document databases also use such concepts to locate related information from huge text archives. Thus concept identification is a key process in leveraging potential knowledge out of the data. Understanding, analyzing and summarizing key concepts from large volumes of text data is non-trivial and crucial in knowledge discovery process.

We address the problem of interpreting statistically generated topics by inducing concepts from large document collection using probabilistic topic modeling. Different from previous studies, we attempt to extract semantically rich “concepts” directly from LDA [1] generated “topics” using a simple framework which is completely unsupervised and easy to implement. Exploration in these dimensions may enrich existing systems in better understanding of text data and can be applied to tasks such as document summarization [14], automated ontology generation [15] etc. We also show that our framework is fit for the task of searching and retrieving documents in large archives and outperforms state of the art probabilistic and phrase based mechanisms.

The main contributions of the paper are summarized as follows:

1. Proposes a novel framework for extracting semantically rich and close to real-world concepts directly from statistically generated topics using a lightweight scoring algorithm.
2. Rigorous experimental comparison of the proposed concept extraction method with state-of-the-art approaches, establishes the effectiveness of the proposed method.
3. Demonstrates the usefulness of the method in improving document retrieval performance in information retrieval tasks.

Organization: The rest of this paper is organized as follows. We review related works in Sect. 2. Section 3 briefly outline the background on probabilistic topic model and LDA algorithm we use in this paper. Section 4 specifies the research objective and a detailed explanation of our proposed topic modeling guided

concept extraction algorithm is presented in Sect. 5. In Sect. 6, we show the usefulness of our proposed method in improving the document retrieval and ranking process. We discuss our experimental setup in Sect. 7, the results and detailed evaluation is described in Sect. 8. Finally we draw conclusions and discuss future work in Sect. 9.

2 Related Work

Topic models such as Latent Dirichlet Allocation (LDA) [1] and Probabilistic Latent Semantic Analysis (PLSA) [3] come up with well established mathematical and statistical model to inspect unstructured text documents for bringing out hidden themes called “topics”. Such topics are probability distributions over words in the vocabulary. In this section we evaluate related researches on the dimension of inferring concepts from text data using topic modeling and critically review those works which are closely similar to our proposed method.

The first notable work which explored beyond the traditional “bag-of-word” approach in topic modeling is the Bigram topic model [16]. In this model each topic word is generated from the distribution of words over a context which is given by a latent topic and the previous word. Later Wang et al. extended the basic Bigram topic model and proposed a new model called Topical n-gram [11] by incorporating a switching variable at each word position to denote the starting of a new n-gram. If this switching variable is not triggered, then the word will be considered as the continuation of a previously identified n-gram. The major shortfall of this model is that a post-processing is required to get the topic of a final word in a n-gram as the topic of the entire n-gram. This is because, in practical, words within an n-gram will not share same topic normally.

Identifying these shortcomings of the Topical n-gram model [11], Lindsay et al. proposed PDLDA (Phrase discovering topic model) [12] which used the hierarchical Pitman-Yor processes [17] for creating topic-word matrix. A topic segmentation model which incorporates word order [13] is later proposed by Jameel et al. Apart from topic detection, these models performs phrase segmentation which is computationally expensive when dealing with large text archives. More recent work in this dimension is reported in 2014 by El-Kishky et al. in which they proposed a topical phrase mining method called TopMine [18]. This framework uses a two step process - one for discovering phrases from text and next for training a traditional LDA model on these phrases and their assumption is that words in the same phrase should be assigned with the same topic.

Other recent work reported is a framework proposed by Yulan He for extracting topical phrases from clinical documents [19]. In this two step work, the author first extracts medical phrases using an off-the-shelf tool and then train a topic model which takes a hierarchy of Pitman-Yor processes [17] as prior for modeling the generation of phrases of arbitrary length. Other notable work in this dimension is the Hierarchical Concept Topic Model (HCTM) [20] and Concept Topic Model (CTM) [22] which incorporates concepts into probabilistic topic models. HCTM employs a manual tagging mechanism to tag concepts in a document and involves assigning concepts to each word in a document. Then the semantic

themes of a document content are revealed as a probability distribution over concepts and combine a hierarchy of human defined concepts with statistical topic models to combine the best of both. The major disadvantage of this work is that the tagging process is cumbersome and cannot be done without human annotators. A cluster based iterative topical phrase mining framework [31] was recently introduced that present a novel framework for topical phrase mining. Their approach treats corpus as a mixture of clusters and each cluster is characterized by documents sharing similar topical distributions. Then this method iteratively performs phrase mining, topical inferring and cluster updating until a satisfactorily final result is obtained. Another notable work in this dimension was introduced by Li and Jin [32]. They proposed a semantic concept latent dirichlet allocation and semantic concept hierarchical dirichlet process based approaches by representing text as meaningful concepts rather than words. The authors implemented the algorithms in discovering new semantic relation between concepts from text documents. The method improved the search quality when compared with other LDA or HDP based approaches. Another very recent work was introduced by Xu et al. [33] that incorporates Wikipedia concepts and categories as prior knowledge into topic models. Their work utilizes entity knowledge, concepts and categories in Wikipedia as prior knowledge into topic models so that it discover more coherent topics. Their method not only modeled the relationship between words and topics, but also utilizes knowledge of concept and category to model semantic relationship between them.

The method proposed in this paper introduces a framework which extracts semantically rich concepts directly from a collection of probabilistically generated topics using Latent Dirichlet Allocation (LDA) [1] algorithm. This method first generates a bag of word bi-grams and tri-grams from the static document corpus and rank them using our new scoring function. We also show that our proposed method outperforms already existing phrase based document retrieval methods for retrieving and ranking relevant documents from a large text archive.

3 Background: Latent Dirichlet Allocation (LDA)

Inspired from previous topic models, Blei et al. introduced a new topic modeling algorithm known as Latent Dirichlet Allocation (LDA) [1]. This model assumes that a document contain multiple topics and such topics are extracted using a Dirichlet Prior process. In the following section, we will briefly describe the underlying principle of LDA [1]. Even though LDA works well on broad ranges of discrete datasets, the text is considered to be a typical example to which the model can be best applied. The process of generating a document with n words by LDA can be described as follows [1]:

1. Choose the number of words, n , according to Poisson Distribution;
2. Choose the distribution over topics, θ , for this document by Dirichlet Distribution;
 - (a) Choose a topic $T^{(i)} \sim \text{Multinomial}(\theta)$
 - (b) Choose a word $W^{(i)}$ from $P(W^{(i)}|T^{(i)}, \beta)$

Thus the marginal distribution of the document can be obtained from the above process as:

$$P(d) = \int_{\theta} \left(\prod_{i=1}^n \sum_{T^{(i)}} P(W^{(i)}|T^{(i)}, \beta) P(T^{(i)}|\theta) \right) P(\theta|\alpha) d\theta \quad (1)$$

where, $P(\theta|\alpha)$ is derived by Dirichlet Distribution parameterized by α , and $P(W^{(i)}|T^{(i)}, \beta)$ is the probability of $W^{(i)}$ under topic $T^{(i)}$ parameterized by β . The parameter α can be viewed as a prior observation counting on the number of times each topic is sampled in a document, before we actually seen any word from that document. The parameter β is a hyperparameter determining the number of times words are sampled from a topic [1], before any word of the corpus is observed. At the end, the probability of the whole corpus D can be derived by taking the product of all documents' marginal probability as:

$$P(D) = \prod_{i=1}^M P(d_i) \quad (2)$$

4 Research Objective

The following are our main research objectives:

1. Introduce the task of topical phrase discovery from unstructured text corpus and its applications in real life scenarios.
2. Propose a framework that uses a lightweight scoring algorithm for extracting concepts from statistically generated topics using Latent Dirichlet Allocation (LDA).
3. Verify experimentally the effectiveness of the method in extracting real-world concepts. We compare our proposed algorithm with state-of-the-art phrase discovery topic models to establish the fitness of our method for concept extraction.
4. Show the usefulness of our approach in phrase based document retrieval task, where given a user query relevant documents are retrieved from a large static text archival.

5 Topic Modeling Guided Concept Extraction

In this section, we present our approach for extracting close to real-world and semantically interpretable concepts from relatively larger static document corpus. Firstly we present our proposed concept scoring function and the topic modeling guided concept extraction algorithm. Secondly we describe how our proposed method can better rank documents in a large text archival and how our method is efficient in retrieving relevant documents. Firstly, we outline our proposed scoring function that we use to score real-world concepts that are latent in large text document corpora. We introduce a scoring function where the score

of a noun-phrase n in a topic t is computed as the probability of a topic over the document multiplied by the frequency of noun-phrase in that document. Notationally,

$$score(n, t) = \sum_{document_d} p(t|d)freq(n, d) \quad (3)$$

where $p(t|d)$ is the scoring for the document-topic pair obtained from the LDA and $freq(n, d)$ is the frequency of noun-phrase n in document d . After calculating score for each noun-phrase, we filter top k noun-phrases according to the score. Now we propose our algorithm, Algorithm 1, which takes a set of topics which are probability distribution over words, generated by Latent Dirichlet Allocation (LDA) [1] and a set of static corpus of documents D . Using an off-the-shelf noun-phrase tagger, the algorithm first tags all the noun-phrases from static corpus D . The frequencies of noun-phrase in documents are then computed. The probability distribution of topic over documents, $p(t|d)$ is computed from LDA output. Then the score of already tagged noun-phrases in documents, $score(n, t)$ calculated as the product of $p(t|d)$ and $freq(n, d)$. Finally, we find top - k noun-phrases according to the score, $score(n, t)$ calculated using Eq. 3.

Algorithm 1. Algorithm for topic modeling guided concept extraction

```

function TM-ExtractConcepts ( $D, T$ );
Input : A static document collection,  $D$  and a topic collection,  $T$ 
Output: top- $k$  phrases for each topic
for each document  $d$  from  $D$  do
  | run through a noun-phrase tagger, and collect noun-phrases
end
for topic  $t$  and noun-phrase  $n$  do
  |  $score(n, t) = \sum_{document_d} p(t|d)freq(n, d)$ 
end
for each topic  $t$ , find the top- $k$  noun-phrases according to the  $score(n, t)$ 

```

6 Topic Modeling Guided Document Retrieval

Given a multi-word query Q , computing the top k relevant documents from a large collection of documents D is a fundamental problem in Information Retrieval (IR). The relevance between a given query and a document d is determined by using similarity functions such as BM25 [24] and such functions generate scores for every document d in the collection D . Then the system reports k top scored documents as the ranked result. Earlier, vector based models were introduced in which documents are represented as a vector of terms, $\vec{d}_j = w_{1j}, w_{2j}, \dots, w_{tj}$ where t is the total number of words in the document and each $w_{1j} > 0$ if and only if the word i is present in document d_j . In this model, the term weights are not binary as in boolean model since vector space model do

not consider the presence or absence of terms. In the same fashion, query is also represented as vector and the similarity between a query vector and a document vector is calculated for measuring the relevance and the same is used as a ranking score. In vector space model, there are two widely used factors for calculating term weights; namely, term frequency (tf) and inverse document frequency (idf). The term weight is computed by multiplying these two factors and widely known as $tf - idf$ measure [23]. Then the similarity measure between the query vector and document vector is calculated using widely used similarity measures such as Okapi BM25 [24] as:

$$score(q, d) = \sum_{i=1}^{|q|} idf(q_i) \cdot \frac{tf(q_i, d) \cdot (k_1 + 1)}{tf(q_i, d) + k_1 \cdot (1 - b + b \cdot \frac{|d|}{avgdl})} \quad (4)$$

where, $tf(q_i, d)$ is q_i 's term frequency in the document d , $|d|$ is the length of the document d in words, and $avgdl$ is the average document length in the text collection from which documents are drawn. When dealing with large collection of documents, calculating tf , idf and then computing the similarity measures between query and document vectors are often difficult.

Here, we propose a new method for retrieving top - k documents from a large collection of text corpus using [concept - topic - document] triplet where concepts are extracted using the scoring function we have introduced in Sect.6. Given a query, we first find relevant topics using the function $score(n, t) = \sum_{document_d} p(t|d) \cdot freq(n, d)$ where $p(t|d)$ is the probability of topic over document and $freq(n, d)$ is the frequency of phrase over document, which are pre-computed. Using topic distribution over documents information given by LDA, we then obtain a set of documents having highest topic probability and present top- k documents as the ranked collection.

Algorithm 2. Algorithm for topic modeling guided document retrieval

```
function TM-RetrieveDocuments( $q$ );
Input : A phrase query  $q$ 
Output: Ranked set of documents
retrieve the significant topics for  $q$  using scoring function,  $score(q, t)$ , where
 $score(q, t) = \sum_{document_d} p(t|d) \cdot freq(q, d)$ 
use  $score(q, t)$ , to get topic - document distribution (given by LDA)
list top -  $k$  documents according to the distribution
```

7 Experimental Setup

7.1 Dataset Description

We use three different datasets in our experimental evaluation - the BBC dataset, StackExchange dataset and a much larger Reuters 21578 dataset. A short description of these datasets are given below.

- **BBC News Dataset:**¹ The general BBC dataset consist of 2225 text documents directly from their website corresponding to stories in five areas such as business, entertainment, politics, sports and technology, from 2004 to 2005.
- **StackExchange Dataset:**² This dataset comprise of anonymized, user-contributed contents on the StackExchange website in different categories. Since our proposed approach uses a crowd-sourcing experiment we have carefully chosen categories in such a way that it matches with the general knowledge of the users participating. Specifically we have chosen academia, android, cooking, gaming and travel categories for this experiment. All post under a single thread is merged in one document and thus the final dataset consists of 2769 documents.
- **Reuters 21578 Dataset:**³ Reuters 21578 dataset is a collection of newswire articles and is popular among data mining communities. The collection was made available for research purposes by Reuters in 1990.

7.2 Experimental Testbed for Topic Induced Concept Extraction

This section describes the experimental setup we have used for our proposed concept extraction experiment. All methods described in this paper were implemented in Python 2.7. The experiments were run on a server configured with AMD Opteron 6376 @ 2.3 GHz/16 core processor and 16 GB of main memory. Firstly, we pre-processed each document for removing common words such as “the”, “and” etc. We then used TextBlob [27] library in Python for tagging noun-phrases from each document. We considered noun-phrases that contain at least two words and for all datasets, we have filtered noun-phrases that are word n -grams where $n \geq 2$. We then modeled topics using LDA algorithm and the score of each noun-phrase is calculated using Eq. 3 and selected top $-k$ noun-phrases for each topic. We use MALLETT⁴ implementation of the LDA model to generate topics and the number of iterations in Gibbs sampling used in this paper is set as 300, as we find that Gibbs sampling usually approaches the target distribution after 300 rounds of iterations. The parameters α and β are set as $\alpha = 50/Z$ and $\beta = 0.01$, respectively.

Baselines - We compare our proposed concept extraction approach with the following baselines. For all the hyper-parameters, we use the default settings and optimizes these parameters every 100 Gibbs sampling iterations.

- **TNG** [11] - The topical n -gram model automatically decides whether to form an n -gram or not by considering its surrounding text. The TNG model provides a systematic way to model topical phrases and simultaneously detect n -grams and topics. The MALLETT implementation of the TNG is used with all the default settings for hyperparameters.

¹ <http://mlg.ucd.ie/datasets/bbc.html>.

² <https://archive.org/details/stackexchange>.

³ <http://www.daviddlewis.com/resources/testcollections/reuters21578/>.

⁴ <http://mallet.cs.umass.edu/>.

- **TopMine** [18] - This algorithm first extracts phrases using a method similar to frequent pattern mining and then train a modified LDA model on the “bag-of phrases” input.

7.3 Experimental Testbed for Document Retrieval

In this section, we describe the experimental setup used for evaluating document retrieval using our proposed framework. We compare our newly devised model to BM25 [24] which is a term-matching based baseline and two phrase based retrieval baselines that uses a flat position index [28] and inverted index [29] as their data structure. For the BM25 setup, we have used the parameters $k_1 = 1$, $k_2 = 0$, $k_3 = 1$ and $b = 0.5$ for initial experiment and later finalized the values of k_1 and b as 1.7 and 0.95 respectively as these values gave optimum retrieval results for the current experiment. We implemented the inverted index and flat position index using Python 2.7 version and used public code libraries at https://www.rosettacode.org/wiki/Inverted_index_Python and <https://github.com/matteobertozzi/blog-code/blob/master/py-inverted-index/> for developing some of the components of the indexes. Our experiments were run on a server configured with AMD Opteron 6376 with 2.3 GHz with 16 core processor and 16 GB of main memory.

8 Result and Evaluation

8.1 Qualitative Evaluation

For measuring closeness of the concepts generated using our proposed method with the real-world concepts, we have conducted a crowd-sourcing experiment which is widely used in text mining and processing tasks to validate and verify the computer generated outputs with human annotated contents. Crowd-sourcing has been applied to tasks such automated question-answering systems [25] and ontology alignment [26]. For this experiment, we have created a web interface that present the users with the concepts extracted by our algorithm and asked them to positively reward every concept which seems to be semantically valid. We used Fleiss Kappa score [30] for the inter-annotator agreement. Measures such as $precision(P)$, $recall(R)$ and $F - measure(F)$ is used for computing the quality of extracted concepts to the human identified concepts. Here, true positive is defined as the number of overlapped concepts between human authored concepts and concepts generated by the algorithm, false positive is the number of extracted concepts that are not truly human authored concepts and false negative is the human authored concepts that are missed by the concept extraction method. Our precision, recall and f-measure values are shown in Table 1 and from the values we draw the following conclusions. We have got precision values ranging from 90.12% to 94.58%. This shows our method extracts quality concepts which are close to real world when compared with the results of a crowd-sourcing experiment (Fig. 1).

Table 1. Comparison of proposed topic modeling guided concept extraction method with Topical N-gram, TopMine baselines. Proposed method outperforms significantly better than all the baselines on all datasets.

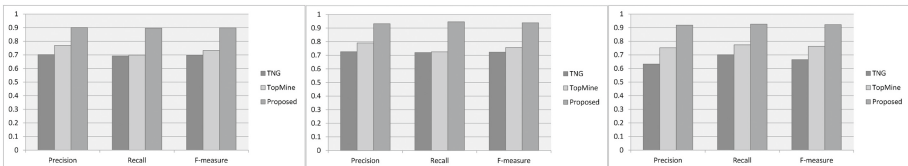
Method	Dataset								
	Reuters			BBC			StackExchange		
	P	R	F	P	R	F	P	R	F
TNG [11]	0.7015	0.6922	0.6968	0.7259	0.7199	0.7228	0.6325	0.7011	0.6650
TopMine [18]	0.7692	0.6988	0.7323	0.7903	0.7254	0.7564	0.7521	0.7741	0.7629
Proposed	0.9012	0.8966	0.8988	0.9318	0.9458	0.9387	0.9187	0.9258	0.9222

Table 2. Comparison of proposed document retrieval method with BM25, flat position index and inverted index baselines.

Method	Dataset								
	Reuters			BBC			StackExchange		
	P	R	F	P	R	F	P	R	F
BM25 [24]	0.6432	0.5920	0.6165	0.5177	0.4852	0.5009	0.6129	0.6233	0.6180
FPI [28]	0.7012	0.6903	0.6957	0.6911	0.7230	0.7066	0.6709	0.6237	0.6464
Inv. idx [29]	0.8126	0.8701	0.8403	0.7752	0.7659	0.7705	0.8129	0.8788	0.8445
Proposed	0.9127	0.9201	0.9163	0.9322	0.9412	0.9366	0.8099	0.8123	0.8110

8.2 Evaluation of Document Retrieval

Here we evaluate our proposed document retrieval method and compares it with the baselines such as [24, 28, 29]. Table 2 shows the performance comparison in terms of precision, recall and f-measure. From the table it is evident that our proposed method significantly outperforms all the three baselines on two out of three datasets we have chosen. The recorded precision values for our proposed method are ranging from 91.27% to 94.12% for BBC and Reuters dataset respectively. For the StackExchange dataset, we have noted that our closest competitor [29] achieved better precision than the proposed method. This is due to comparatively less number of concepts extracted for different categories we



(a) Precision, Recall and F-measure comparison on Reuters dataset (b) Precision, Recall and F-measure comparison on BBC dataset (c) Precision, Recall and F-measure comparison on StackExchange dataset

Fig. 1. Precision, Recall and F-measure values of concept extraction when compared with baseline methods on Reuters, BBC and StackExchange datasets

have considered for the StackExchange dataset such as academia, android and cooking. We found that the noun-phrase tagger was unable to tag phrases from those less popular categories (Fig. 2).

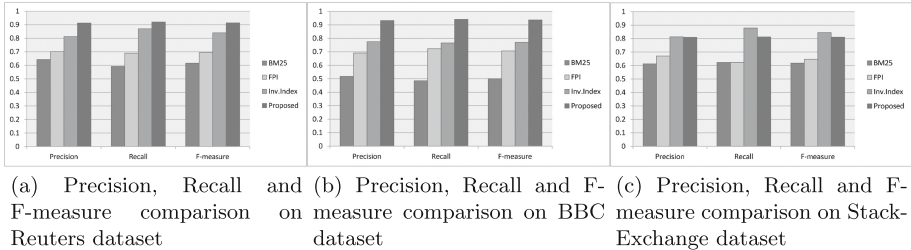


Fig. 2. Precision, Recall and F-measure values obtained for document retrieval when compared with baseline methods on Reuters, BBC and StackExchange datasets

9 Conclusions and Future Work

In this paper, we proposed a novel framework for leveraging semantically rich concepts from statistically computed topics using a widely used LDA algorithm. Rigorous experiments with three real-world datasets show that our proposed method produce close to the real-world concepts which we verifies with a crowdsourcing experiment in terms of precision, recall and f-measure. We also demonstrate the usage of our framework on document retrieval and ranking which is a well studied area in information retrieval. The experimental results show that our method can significantly outperform phrase based document retrieval baselines. There are further scopes for future work. By constantly improving the quality of concepts by tuning the parameters of our proposed algorithm, the extracted concepts can be used for building ontologies by exploiting the relations between them. Concept hierarchy learning can also be done using this and we plan to extend our framework further in these dimensions.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003)
2. Hofmann, T.: Probabilistic latent semantic indexing. In: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 50–57. ACM (1999)
3. Hofmann, T.: Probabilistic latent semantic analysis. In: *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pp. 289–296. Morgan Kaufmann Publishers Inc. (1999)
4. Chang, J., Boyd-Graber, J.L., Gerrish, S., Wang, C., Blei, D.M.: Reading tea leaves: how humans interpret topic models. In: *NIPS*, vol. 31, pp. 1–9 (2009)
5. Arvanitis, A., Wiley, M. T., Hristidis, V.: Efficient Concept-based Document Ranking. In: *EDBT*, pp. 403–414 (2014)

6. Egozi, O., Gabrilovich, E., Markovitch, S.: Concept-based feature generation and selection for information retrieval. In: AAAI, pp. 1132–1137 (2008)
7. Celikyilmaz, A., Hakkani-Tr, D.: Concept-based classification for multi-document summarization. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5540–5543 (2011)
8. Cambria, E.: An introduction to concept-level sentiment analysis. In: MICAI, vol. 2, pp. 478–483 (2013)
9. Asharaf, S., Anoop, V.S., Afzal, A.L.: A framework for meaning aware product discovery in e-commerce. In: Encyclopedia of e-Commerce Development, Implementation, and Management, pp. 1386–1398. IGI Global (2016)
10. Anoop, V.S., Asharaf, S.: A topic modeling guided approach for semantic knowledge discovery in e-commerce. *Int. J. Interact. Multimedia Artif. Intell.* **4**, 1–8 (2017)
11. Wang, X., McCallum, A., Wei, X.: Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In: Proceedings of Seventh IEEE International Conference on Data Mining, ICDM 2007, pp. 697–702 (2007)
12. Lindsey, R.V., Headden III, W.P., Stipicevic, M.J.: A phrase-discovering topic model using hierarchical pitman-yor processes. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 214–222 (2012)
13. Jameel, S., Lam, W.: An unsupervised topic segmentation model incorporating word order. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 203–212 (2013)
14. Yang, G., Wen, D., Chen, N.S., Sutinen, E.: A novel contextual topic model for multi-document summarization. *Exp. Syst. Appl.* **42**(3), 1340–1352 (2015)
15. Sleeman, J., Finin, T., Joshi, A.: Topic modeling for RDF graphs. In: LD4IE@ ISWC, pp. 48–62 (2015)
16. Wallach, H.M.: Topic modeling: beyond bag-of-words. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 977–984 (2006)
17. Pitman, J., Yor, M.: The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25**(2), 855–900 (1997)
18. El-Kishky, A., Song, Y., Wang, C., Voss, C.R., Han, J.: Scalable topical phrase mining from text corpora. *Proc. VLDB Endow.* **8**(3), 305–316 (2014)
19. He, Y.: Extracting topical phrases from clinical documents. In: AAAI, pp. 2957–2963 (2016)
20. Chemudugunta, C., Smyth, P., Steyvers, M.: Combining concept hierarchies and statistical topic models. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, pp. 1469–1470 (2008)
21. Anoop, V.S., Asharaf, S., Deepak, P.: Unsupervised concept hierarchy learning: a topic modeling guided approach. *Procedia Comput. Sci.* **89**, 386–394 (2016)
22. Chemudugunta, C., Holloway, A., Smyth, P., Steyvers, M.: Modeling documents by combining semantic concepts with unsupervised statistical learning. In: International Semantic Web Conference, pp. 229–244 (2008)
23. Ramos, J.: Using tf-idf to determine word relevance in document queries. In: Proceedings of the First Instructional Conference on Machine Learning (2003)
24. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., Gatford, M.: Okapi at TREC-3, p. 109. NIST Special Publication, Gaithersburg (1995)
25. Mrozinski, J., Whittaker, E., Furu, S.: Collecting a why-question corpus for development and evaluation of an automatic QA-system. In: 46th Annual Meeting of the Association of Computational Linguistics: Human Language Technologies, pp. 443–451 (2008)

26. Sarasua, C., Simperl, E., Noy, N.F.: Crowdmap: Crowdsourcing ontology alignment with microtasks. In: International Semantic Web Conference, pp. 525–541 (2012)
27. Loria, S.: TextBlob: simplified text processing (2014)
28. Shan, D., Zhao, W.X., He, J., Yan, R., Yan, H., Li, X.: Efficient phrase querying with flat position index. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, pp. 2001–2004 (2011)
29. Patil, M., Thankachan, S.V., Shah, R., Hon, W.K., Vitter, J.S., Chandrasekaran, S.: Inverted indexes for phrases and strings. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 555–564 (2011)
30. Fleiss, J.L., Cohen, J.: The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ. Psychol. Measur.* **33**(3), 613–619 (1973)
31. Li, B., Wang, B., Zhou, R., Yang, X., Liu, C.: CITPM: a cluster-based iterative topical phrase mining framework. In: International Conference on Database Systems for Advanced Applications, pp. 197–213 (2016)
32. Li, X., Jin, W.: Cross-document knowledge discovery using semantic concept topic model. In: 15th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 108–114 (2016)
33. Xu, K., Qi, G., Huang, J., Wu, T.: Incorporating Wikipedia concepts and categories as prior knowledge into topic models. *Intell. Data Anal.* **21**(2), 443–461 (2017)

Markov Chain Monte Carlo Methods and Evolutionary Algorithms for Automatic Feature Selection from Legal Documents

S. Pudaruth¹(✉), K.M.S. Soyjaudah², and R.P. Gunputh³

¹ Department of Ocean Engineering and ICT, Faculty of Ocean Studies,
University of Mauritius, Moka, Mauritius

s.pudaruth@uom.ac.mu

² Department of Electrical and Electronic Engineering, Faculty of Engineering,
University of Mauritius, Moka, Mauritius

s.soyjaudah@uom.ac.mu

³ Department of Law, Faculty of Law and Management, University of Mauritius,
Moka, Mauritius

rpgunput@uom.ac.mu

Abstract. In this paper, we present three different approaches for feature selection, starting from a naïve Markov Chain Monte Carlo random walk algorithm to more refined methods like simulated annealing and genetic algorithms. It is typical for textual data to have thousands of dimensions in their feature space which makes feature selection a crucial phase before the final classification. Classification of legal documents into eight categories was performed via a simple document similarity measure based on term frequency and the nearest neighbour concept. With an average success rate of 76.4%, the random walk algorithm not only performed better than the simulated annealing and genetic algorithms but also matched the accuracy of support vector machines. Although these methods have commonly been used for selecting appropriate features in other fields, their use in text categorisation have not been satisfactorily investigated. And, to our knowledge, this is the first work which investigates their use in the legal domain. This generic text classification framework can further be enhanced by using an active learning methodology for the selection of training samples rather than following a passive learning approach.

Keywords: Monte Carlo · Random walk · Genetic algorithm · Simulated annealing · Legal text categorisation · Court judgements

1 Introduction

Text categorisation or text classification is a sub-branch of text mining and natural language processing technique which attempts to assign documents to specific categories. Text categorisation is important in many applications such as the automated generation of metadata in document retrieval systems, question answer systems and search engines (Sahin 2007). Text documents are usually represented using the bag-of-words approach and vector space model. In this representation, a document or a

document set can have thousands of vectors where each vector is usually a tuple consisting of a word and its term frequency. Because of the very high dimensions inherent in textual data, feature selection is one of the key steps in text categorisation.

Given a set of n words, it is still possible to generate $2^n - 1$ subsets of 1 to n words and $\frac{n!}{r!(n-r)!}$ possible subsets of size r . Thus, there are more than 47 trillion ways of choosing 30 words from a bag of 50 words. So, for a given document set, it is usually not possible to consider all possible combinations of terms in order to find the best one. Therefore, when no exact formulation of a problem is possible, it is necessary to resort to approximations. In text mining, dimension reduction is usually achieved by converting all text to lowercase, removal of stopwords, stemming and lemmatisation. However, this is often not sufficient and the dimension often remains unusually large.

In this paper, we investigate three different methods which are all based on the principles of Markov chain Monte Carlo (MCMC) methods. These Monte Carlo techniques, due to their general applicability, has been used in a large variety of domains especially for studies involving simulations in physics, biology and computer science (Browne et al. 2012). Andrieu et al. (2003) wrote an interesting paper which bridges the gap between the Monte Carlo methods and traditional machine learning techniques. The basic Monte Carlo techniques and its various variance-reduction variants and their applications are described in detail. A theoretical foundation of the different techniques are also provided.

The use of simulated annealing and genetic algorithms as feature selection methods in the text mining field are scarce compared to other fields but not new. Genetic algorithms have been used in information retrieval systems by Gordon (1988) and Chen and Kim (1994). Although simulated annealing (Metropolis et al. 1953) is an older technique than genetic algorithm (Holland 1975), its use as a feature selection method in text classification is relatively recent (Wang et al. 2006; Yang et al. 2007).

In this work, inspired by the principles of Monte Carlo, we are able to show that by repeatedly determining the category of a document from a sample of features selected through a random walk or differential evolution, we are able to deduce the category of a court judgement to a higher accuracy using only a very small fraction of the total number of features. The use of the Monte Carlo approach allows a micro-local analysis to be performed on a high-dimensional problem. We believe that this new technique can become a tool of general applicability in the field of text classification.

The performance of machine learning classifiers hinges on the availability of large amounts of training data. Manual classification of court cases requires highly skilled professionals and is a time-consuming and costly undertaking. Through passive selection, we divided our dataset into several partitions of training set (10–90%) and testing set (90–10%). In the future, we intend to use the principles of active learning to optimise the choice of the training samples (Roy and McCallum 2001; Figueroa et al. 2012). Although support vector machines performs better on a 90–10 split, our Monte Carlo-based classification is able to achieve higher accuracies when the training set is less than 80% of the full dataset.

The remainder of this paper is organised as follows. Section 2 discusses on prior research on Markov Chain Monte Carlo methods, simulated annealing and genetic algorithms. Section 3 describes in detail how these methods have been adapted for

feature selection and text classification in our system. The experimental results and discussions are provided in Sect. 4. And finally, the conclusion is offered in Sect. 5 wherein some potential avenues for future research is indicated.

2 Related Works

For a modern and comprehensive literature review on the field of text document classification, the reader is referred to Khan et al. (2010). The paper starts with describing state-of-the-art applications of text mining and document classification. The pre-processing steps such as feature extraction, feature selection, dimensionality reduction and document representation are also explained. Machine learning techniques such as kNN, naïve Bayes, decision trees, support vector machines, artificial neural networks, fuzzy logic, genetic algorithms, classification rules and hybrid techniques have also been well tackled in reasonable depth.

The Monte Carlo method find its roots in the field of statistics in which random numbers are used to perform simulation (Liang and Wong 2000; Goncharov et al. 2007; Houghton et al. 2014). This method is often used in problems involving very high dimensions, for example, in the Travelling Salesman Problem where there is a very large number of potential solutions. The problem becomes rapidly intractable because of the exponential rise in the number of paths for each new node that is added to the existing network (Buxey 1979; Martin et al. 1991).

Monte Carlo simulation approaches has found wide applications in computer games (Browne et al. 2012). An algorithm known as the Monte-Carlo Tree Search Solver (MCTS-Solver) has been successfully applied in the popular game of Line of Action in order to find better strategies to play the game (Winands et al. 2008). It was shown that a program using the MCTS-Solver was able to win in 65% of matches.

An interesting application of the Monte Carlo Search Framework has been described in Branavan et al. (2012). In particular, they used a Monte Carlo approach to provide textual information retrieved from a manual to feed a game agent. This game-playing agent was able to outperform the uninformed and untrained agent (default AI) in the game of Civilization by a factor of 34%.

Another modern area where Monte Carlo simulations are being heavily used is in the field of bioinformatics. Ebbert et al. (2011) have used the Monte Carlo approach to generate samples for characterising uncertainty in multi-variate assays. This uncertainly information is very important for clinicians in order to properly classify different types of tumours. Diaconis (2009) has used the principles of Markov Chain Monte Carlo in order to decrypt messages. More recently, Monte Carlo methods have been applied in multi-label classification (Read et al. 2014).

The concept of an evolutionary Monte Carlo method is not new. Previously, it has been used in the scientific, engineering and mathematical filed for the estimation of various parameter values (Smith and Hussain 2012) and uncertainties in optimisation problems (Ter Braak 2006; Wu et al. 2006; Xiao 2007).

One of the earliest works which applied a genetic algorithm (GA) for the extraction of correlated terms was done by Desjardins et al. (2005). However, they concluded that the co-occurrences found by the GA did not improve the retrieval accuracy and more

work was required to understand the cognitive factors which would improve the relevancy of retrieved results.

Gavrilis et al. (2005) have used a GA for feature selection on 650 PUBMED abstracts spread into 5 categories. Using an SVM classifier, they achieved 85% accuracy using 20 features only. In another similar study on spam emails, they report an accuracy of 97%, again using only 20 features (Gavrilis et al. 2006).

A comparison was made between tf-idf (term frequency-inverse document frequency) and genetic algorithms by Khalessizadeh et al. (2006) for the identification of document topics in relatively short Persian texts. While precision values were very similar for both methods, GA did better than tf-idf on recall for all document sizes.

Pietramala et al. (2008) proposed the Olex-GA algorithm which assigns positive and negative rules to each feature (gene) in a chromosome. Their approach performed better on the OHSUMED dataset when compared with techniques such as naïve Bayes, C4.5, Ripper and Support Vector Machines. However, for the Reuters-21578 dataset, SVM did significantly better than Olex-GA.

Song and Park (2009) have used a genetic algorithm to find the optimal number of clusters in the Reuters-21578 dataset. The number of terms were reduced using latent semantic indexing (LSI) before the GA was applied. They were able to show that their algorithm performs better than previous methods. However, their study was based on only 1000 documents from 5 categories. Liu and Fu (2012) used an elitist GA to classify 100 web pages into 5 categories and obtained better performance than when using SVM with default parameters.

Chen et al. (2013) proposed a novel text classification procedure based on the chaos optimization theory and genetic algorithm. They tested their approach on the Reuters-21578 dataset and they were able to demonstrate that the algorithm requires a small feature set in order to provide comparable recall performance compared with earlier higher-dimensional techniques. Using GATE (Cunningham and Tablan 2002) and Weka (Hall et al. 2009), Rogers (2013) implemented a single pipeline for feature selection using genetic algorithms and classification several machine learning classifiers. Pavlyshenko (2014) have used a genetic algorithm to determine the optimal subset of keywords which could be used to identify an authors of English fiction texts.

A good introduction to feature selection in text mining using simulated annealing can be found in Bagheri et al. (2014). They demonstrated that their proposed approach delivered similar performance to chi-squared when tested on a Persian dataset consisting of 7 categories. There were about 800 documents in each category. Zhu et al. (2015) further showed that an improved simulated annealing algorithm (SAA) can select better features than information gain (IG), mutual information (MI) and chi-squared (CHI).

Moshki et al. (2015) tested an extended version of the simulated algorithm (SAGRASP) on a diverse set of data and showed that it was able to do better than FCGRASP (Bermejo et al. 2011). Zhu et al. (2009) successfully combined the genetic algorithm and simulated annealing (SA) to extract protein sequences using OpenMP. They reported that their proposed solution did not get trapped in local maxima and could find global optima faster. Two decades earlier, Esbensen and Mazumder (1994) used a mixture of SA and GA, which they called SAGA, to determine an optimal placement for macro-cells.

3 Description of Algorithms

3.1 Markov Chain Monte Carlo Random Walk (MCMCRW)

A random walk is a random or stochastic process that may consist of a series of random steps (Pemantle 2007; Samad 2013). The principles of random walk has found wide applications in different fields of computer science such as the analysis of computer networks (Zhong et al. 2008), computer security (Zhou 2016), bioinformatics (Draminski et al. 2008) and text classification (Hassan et al. 2007).

In our system, a random walk consists of a sequence of similar operations in which a subset of k elements are randomly selected (with replacement) from a list of n elements from each of the m categories, p number of times. In computer science, this is known as a Markov Chain Monte Carlo (MCMC) process. Each element is a word and these $m * n$ elements (mainlist) are initially selected using term frequency. We have two sets of data: the training set and the testing set. The training set is only used for the extraction of representative elements for each category. We do not use a wrapper-style classifier (Jovic et al. 2015) to measure the classifier accuracy, instead we use a naïve classifier based on term frequency. Each of the m sublists is compared to every document in the testing set and the sum of all the k words is computed. The sublist with the highest score is taken as the predicted category. This classifier can be considered as a simplified version of the k -nearest neighbour classifier. A simple majority voting is then carried out on the results obtained after the p iterations.

3.2 Boosted Simulated Annealing (BSA)

The basic principles in the simulated annealing algorithm was described by Metropolis et al. (1953). Their aim was to simulate the movement of atoms at a finite temperature. In 1970, Hastings showed how this method could be generalised to solve problems in statistics. Kirkpatrick et al. (1983) took up the same basic ideas but added the concepts of high and low temperatures and compared the algorithm to the annealing process in metals, from which the algorithm got its name. It is only very recently that the simulated annealing algorithm has been used for feature selection in the text classification field (Wang et al. 2006; Yang et al. 2007; Bagheri et al. 2014; Zhu et al. 2015; Moshki et al. 2015).

In our system, we implemented the standard simulated annealing algorithm but is boosted with a good initial sample which is produced by random sampling through a random walk of 100 steps. The selection of elements is similar to MCMCRW. However, in BSA, the next step is not independent on the current one. The next list is generated by replacing t elements in the current best list by t other elements from each category (sublist) from the mainlist. The number t is a temperature variable which decreases steadily to 0 from the first iteration until the last one. If the accuracy increases after this small change, the best list is updated. However, even if the accuracy decreases by x percent, we still consider this new list as the current one. If the accuracy decreases by x percent or more, the current list is discarded and a new one is generated. This entire process is repeated q times. Our algorithm is also stateful in that it has a memory to store the best list from any of the q iterations.

3.3 Genetic Algorithms

A genetic algorithm (GA) is often described as a meta-heuristic algorithm which emulates the Darwinian's theory of natural evolution through the biological processes of selection, mating and mutation (Mitchell 1998). Since its formulation in 1970 by Holland, GAs has found wide applications, not only in scientific areas but also in business applications. Thomas and Sycara (2002) have used GAs for predicting stock prices while Borg (2009) has used GAs for the automatic extraction of definitions. In combination with other methods, Waad et al. (2014) used a genetic algorithm to select the best features to assess the credit worthiness of a potential client. A recent and novel application of GA is in the detection of errors in SQL instructions (Moncao et al. 2013). Also, as stated earlier, GAs have also been used in the information retrieval domain since more than two decades (Atkinson-Abutridy et al. 2004; Al-Maqaleh et al. 2012).

In our system, we maintain a population of r mainlist. In GA's jargon, a mainlist can be considered as a chromosome. A mainlist is a list of sublists and each sublist contains the most frequent terms in one category of documents. As mentioned earlier, each term is a word (gene). A fitness score (classification accuracy) is calculated for each of these r lists. The best b lists (chromosomes) are selected in each iteration for crossover and mutation. Each of these b lists are randomly paired with each other (without duplication) to generate $\frac{b}{2}$ pairs and c elements are then chosen randomly from one member in each pair and are swapped. This is the crossover operation whereby b new lists are created and the previous b lists are discarded*. Our crossover operation is slightly different from previous approaches that use fixed locations for genes in that we do not exchange a segment of the chromosome, instead, we exchange genes selected randomly from anywhere in the chromosome. The rationale for this heuristic approach is that the genes (words) are independent and this reduces the likelihood of collisions. The next operation is mutation which is achieved by the substitution of d elements in each list by new elements from the mainlists. The remaining $r-b$ lists are rejected and new ones are generated randomly in order to keep the size of the population constant. These processes are repeated q times. The best list from each of these q iterations are stored in a separate memory*.

4 Experiments and Settings

4.1 Experimental Corpus

Our dataset consists of 294 judgements which were delivered in the Supreme Court of the Republic of Mauritius in the year 2013. The cases have previously been classified manually into eight categories: homicide, road traffic offences, drugs, other criminal offences, company law, labour law, land law and contract law. It is the same dataset that we have used previously in an earlier work (Pudaruth et al. 2016).

4.2 Document Pre-processing

Because textual data is inherently noisy, it is important to filter out those elements that would negatively impact on the performance of classifiers. Thus, all the data was first converted to lowercase after which all digits, symbols, short words and stopwords were filtered out. Besides the common English stopwords, the list also included words from the legal domain such as case, act, section, law, court, appellant, respondent, judge and many others. These words tend to occur in almost all judgements and their frequencies are also very high. This is an interesting issue because the same words could have been strong differentiators if our objective was to look for legal documents in a mixture of documents from other domains. With the help of Wordnet (2017), all non-English words and all verbs were removed from the dataset. All the pre-processing steps taken together reduced the feature set from 6048 to 2485 words.

4.3 Experimental Environment and Algorithmic Parameters

The documents are kept in a parent folder with eight directories. Each directory contains the files for one specific category whereby each judgement is stored as a separate textfile. The software for document pre-processing and feature selection have been implemented in Python 2.7.12 (Spyder 3.0.0) from the Anaconda distribution. The scikit-learn library for Python has been used for the machine learning part. A computer running the 64-bit Windows 7 Professional N (SP1) operating system has been used in this study. The processor is an Intel(R) Core(TM) i5-4200 M CPU running at 2.5 GHz on 8.00 GB of RAM on a hard disk of 650 GB.

The number of iterations for each of the 3 Monte Carlo methods was 100. A sample consisted of 30 (k) elements drawn from a larger set of 100 (n) most frequent words from each of the 8 (m) categories. The mutation rate for both the simulated annealing and genetic algorithms was 20%. This means that for every 30 elements, 6 elements were exchanged. The initial temperature for SA was set at 10 (for every sublist) and this was reduced by 0.1 after every iteration until it reached a minimum value of 1. The SA algorithm was allowed to accept solutions which was 5% (x) worse than the current one in an attempt to avoid local maxima. The crossover rate for the GA was also set at 20%. The parameters were chosen empirically, after conducting a large set of experiments and observing their impact on the accuracy. All the algorithms were run 5 times on each split percentage and an average was made.

4.4 Experimental Results

In this sub-section, we present the detailed results to demonstrate the effectiveness of the Markov Chain Monte Carlo (MCMC) Random Walk algorithm compared to simulated annealing (SA), genetic annealing (GA) and support vector machines (SVM). The dataset has been split into nine different sets of training and testing data, as shown in Table 1, with a view to understand the influence of different training sizes on feature selection and classification accuracy. A comparison with SVM is also provided.

Table 1. Details of cases dataset

Categories	Code	No. of Cases
Company Law	Comp	22
Contract Law	Cont	56
Other Criminal Offences	Crim	48
Drugs	Drugs	44
Homicide	Homi	14
Labour Law	Labo	17
Land Law	Land	55
Road Traffic Offences	Road	38

In general, classification accuracy increases when the size of the training set increases, as shown in Table 2. For all training sizes, MCMCRW and SVM are more effective than SBA and GA. The best accuracy of 83% is obtained by MCMCRW at 70% of the training set and by SVM at 90% of the training set. When the training size is 20% or less, all the three algorithms (MCMCRW, BSA and GA) does better than SVM (with default settings and parameters). When the training size is between 40 and 70%, only MCMCRW is able to outperform SVM. The results illustrate that our random walk algorithm can produce satisfactory results even with low amount of training data. In the legal field where it is very costly and time-consuming to produce annotated data as this has to be done by highly trained professionals, the random walk algorithm only requires a few relevant documents for training for each category to deliver acceptable results.

Table 2. Classification accuracy v/s training/testing size

Training Set (%)	10	20	30	40	50	60	70	80	90	Average
Testing Set (%)	90	80	70	60	50	40	30	20	10	
MCMC Random Walk	68	76	78	70	76	80	83	78	79	76.4
Boosted Simulated Annealing (BSA)	64	63	63	61	66	65	69	74	76	66.8
Genetic Algorithm (GA)	62	66	64	60	66	64	72	73	79	67.3
Support Vector Machines (SVM)	53	54	63	67	73	76	75	78	83	69.1

Table 3 shows the detailed results for one run on a split of 70/30, in which there were 201 cases in the training set and 93 cases in the testing set. The overall accuracy for this run was 84%. Accuracy is defined as the number of correctly classified documents over the total number of documents in the testing set. The *Road traffic offences* category has a perfect recall, which means that we have been able to retrieve all instances of this category from the testing set, while the *Contract* category has the lowest recall because 3 of its cases have been incorrectly retrieved as a *Labour* case and another two as a *Land* case. However, these misclassifications are quite comprehensible as these 3 categories share many terms in common as they all deal with contractual issues.

Table 3. Confusion matrix

Category	Comp	Cont	Crim	Drug	Homi	Labo	Land	Road	Total	Recall
Comp	6	0	0	0	0	0	1	0	7	0.86
Cont	0	12	0	0	0	3	2	0	17	0.71
Crim	0	0	13	1	0	0	0	1	15	0.87
Drug	0	0	2	12	0	0	0	0	14	0.86
Homi	0	0	0	1	4	0	0	0	5	0.80
Labo	1	0	0	0	0	5	0	0	6	0.83
Land	2	1	0	0	0	0	14	0	17	0.82
Road	0	0	0	0	0	0	0	12	12	1.00
Total	9	13	15	14	4	8	17	13	93	0.82
Precision	0.67	0.92	0.87	0.86	1.00	0.63	0.82	0.92	0.84	

The *Homicide* category has the highest precision followed closely by *Road traffic offences* and *Contract*. The *Labour* and *Company* categories have the lowest precision values. Nine documents have been returned as belonging to the *Company* category, however, only six of them are actually company law cases. One document belongs to the labour law category while another two belong to the land law category. Again, we see that there is some overlap in features between the *Contract*, *Company* and *Labour* categories. Some additional work will be necessary in order to reduce this mix-up.

The Olex-GA system proposed by Pietramala (Pietramala et al. 2008) did slightly better than SVM on the OHSUMED dataset but less well on the Reuters-21578 dataset. The SA algorithm proposed by Yang et al. (2007) performed slightly better than kNN on some settings. Wang et al. (2006) reported a similar result. However, it is always very difficult to offer a fair comparison when comparing GAs and SAs with machine learning classifiers. The variation in the total number of documents in the dataset, the number of classes, the size of the documents, the number of training & testing samples, the nature & complexity of the documents, the multitude of parameters used in the evolutionary algorithms and in the classifiers all lead to a very difficult comparison between the various studies.

5 Conclusions

This paper presents three different feature selection methods and a general text categorisation framework. After an extensive empirical evaluation with a set of 294 Supreme Court judgements spread into eight areas of law, we found that the simulated annealing and genetic algorithms, with an average classification accuracy of about 67%, did less well than the conceptually simpler random walk while the latter's performance was on average better than support vector machines. The idea of using evolutionary operations for the selection of suitable features is not new, however, full-fledged implementation of these techniques in the domain of text classification is relatively recent. The random walk algorithm is very robust as it does not fluctuate as much as machine learning classifiers do when the number of training samples is

reduced. The added benefit of this system is its simplicity and understandability. It is very easy for a user to improve the quality of the process by adding new training samples or new filters. In future work, we intend to choose the training instances using an active learning technique instead of passive learning. Combining the strengths of each of these feature selection methods into a single algorithm is also a potential avenue for further research.

References

- Al-Maqaleh, B.M., Shahbazkia, H.: A genetic algorithm for discovering classification rules in data mining. *Int. J. Comput. Appl.* **41**(18), 40–44 (2012)
- Andrieu, C., de Freitas, N., Doucet, A., Jordan, M.I.: An introduction to MCMC for machine learning. *Mach. Learn.* **50**, 5–43 (2003)
- Atkinson-Abutridy, J., Mellish, C., Aitken, S.: Combining information extraction with genetic algorithms for text mining. *IEEE Intell. Syst.* **19**(3), 22–30 (2004)
- Bagheri, A., Saraee, M., Nadi, S.: PSA: a hybrid feature selection approach for Persian text classification. *J. Comput. Secur.* **1**(4), 261–272 (2014)
- Bermejo, P., Gamez, J.A., Puerta, J.M.: A GRASP algorithm for fast hybrid filter-wrapper feature subset selection in high-dimensional datasets. *Pattern Recogn. Lett.* **32**(5), 701–711 (2011)
- Borg, C.: Automatic Definition Extraction using Evolutionary Algorithms. Thesis (MSc), University of Malta, Malta (2009)
- Branavan, S.R.K., Silver, D., Barzilay, R.: Learning to win by reading manuals in a Monte Carlo framework. *J. Artif. Intell. Res.* **43**, 661–704 (2012)
- Browne, C., Powley, E., Whitehouse, D., Lucas, S., Cowling, P.I., Rohlfshagen, P., Tavener, S., Perez, D., Samothrakis, S., Colton, S.: A survey of Monte Carlo tree search methods. *IEEE Trans. Comput. Intell. AI Games* **4**(1), 1–43 (2012)
- Buxey, G.M.: The vehicle scheduling problem and Monte Carlo simulation. *J. Oper. Res. Soc.* **30**(6), 563–573 (1979)
- Chen, H., Kim, J.: GANNET: a machine learning approach to document retrieval. *J. Manag. Inf. Syst.* **11**(3), 7–41 (1994)
- Chen, H., Jiang, W., Li, C., Li, R.: A heuristic feature selection approach for text categorization by using chaos optimization and genetic algorithm. *Math. Problems Eng.* 2013, Article ID: 524017
- Cunningham, M., Tablan, B.: GATE: a framework and graphical development environment for robust NLP Tools and applications. In: *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL 2002)*, 7–12 July 2002, Philadelphia, Pennsylvania (2002)
- Desjardins, G., Godin, R., Proulx, R.: A genetic algorithm for text mining. *WIT Trans. Inf. Commun. Technol.* **35**, 133–142 (2005)
- Diaconis, P.: The Markov chain Monte Carlo revolution. *Bull. Am. Math. Soc.* **46**, 179–205 (2009)
- Draminski, M., Rada-Iglesias, A., Enroth, S., Wadelius, C., Koronacki, J., Komorowski, J.: Monte Carlo feature selection for supervised classification. *Bioinformatics* **24**(1), 110–117 (2008)
- Ebbert, M.T.W., Bastien, R.R.L., Boucher, K.M., Martin, M., Carrasco, E., Caballero, R., Stijleman, I.J., Bernard, P.S., Facelli, J.C.: Characterization of uncertainty in the classification of multivariate assays: application to PAM50 centroid-based genomic predictors for breast cancer treatment plans. *J. Clin. Bioinform.* **1**, 37 (2011)

- Esbensen, H., Mazumder, P.: SAGA: a unification of the genetic algorithm with simulated annealing and its application to macro-cell placement. In: Proceedings of the 7th International Conference on VLSI Design, Calcutta, India, 5–8 January 1994, pp. 211–214 (1994)
- Figueroa, R.L., Zeng-Treitler, Q., Ngo, L.H., Goryachev, S., Wiechmann, E.P.: Active learning for clinical text classification: is it better than random sampling? *J. Am. Med. Inform. Assoc.* **19**(5), 809–816 (2012)
- Gavrilis, D., Tsoulos, I.G., Dermatas, E.: Stochastic classification of scientific abstracts. In: Proceedings of the 6th Speech and Computer Conference, Patras, Greece (2005)
- Gavrilis, D., Tsoulos, I.G., Dermatas, E.: Neural recognition and genetic features selection for robust detection of E-mail spam. *Adv. Artif. Intell.* **3955**, 498–501 (2006)
- Goncharov, Y., Okten, G., Shah, M.: Computation of the endogenous mortgage rates with randomized quasi-Monte Carlo simulations. *Math. Comput. Model.* **46**(3–4), 459–481 (2007)
- Gordon, M.: Probabilistic and genetic algorithms for document retrieval. *Commun. ACM* **31**(10), 1208–1218 (1988)
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. *SIGKDD Explor.* **11**(1), 10–18 (2009)
- Hassan, S., Mihalcea, R., Banea, C.: Random walk term weighting for improved text classification. *Int. J. Semant. Comput.* **1**(4), 421–439 (2007)
- Hastings, W.K.: Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**(1), 97–109 (1970)
- Holland, J.H.: *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Michigan (1975)
- Houghton, J., Siegel, M., Wirsch, A., Moulton, A., Madnick, S., Goldsmith, D.: A survey of methods for data inclusion in system dynamics models: methods, tools and applications. Massachusetts Institute of Technology, Cambridge, Working Paper CISL# 2013-03 (2014)
- Jovic, A., Brkic, K., Bogunovic, N.: A review of feature selection methods with applications. In: Proceedings of the 38th IEEE International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2015), Opatija, Croatia, 25–29 May 2015, pp. 1200–1205 (2015)
- Khalessizadeh, S.M., Zafarian, R., Nasser, S.H., Ardil, E.: Genetic mining: using genetic algorithm for topic based on concept distribution. In: Proceedings of the World Academy of Science, Engineering and Technology (2006)
- Khan, A., Baharudin, B., Lee, L., Khan, K.: A review of machine learning algorithms for text documents classification. *J. Adv. Inf. Technol.* **1**(1), 4–20 (2010)
- Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Optimization by simulated annealing. *Science* **220** (4598), 671–680 (1983)
- Liang, F., Wong, W.H.: Evolutionary Monte Carlo: applications to Cp model sampling and change point problem. *Stat. Sin.* **10**, 317–342 (2000)
- Liu, X., Fu, H.: A hybrid algorithm for text classification problem. *Electrical review*, R. 88 NR 1b (2012)
- Martin, O., Otto, S.W., Felten, E.W.: Large-step Markov chains for the travelling salesman problem, p. 16. CSETech, Paper (1991)
- Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**(6), 1087–1092 (1953)
- Mitchell, M.: *An Introduction to Genetic Algorithms*. MIT Press, Cambridge (1998)
- Moncao, A.C.L., Camilo-JR, C.G., Queiroz, L.T., Rodrigues, C.L., Leitao-JR, P.S., Vincenzi, A. M.R.: Applying genetic algorithms to data selection for SQL mutation analysis. In: Proceedings of the 15th Annual Conference on Genetic and Evolutionary Computation (GECCO 2013), Amsterdam, The Netherlands, 7–10 July 2013, pp. 207–208 (2013)

- Moshki, M., Kabiri, P., Mohebalhojeh, A.: Scalable feature selection in high-dimensional data based on GRASP. *Appl. Artif. Intell.* **29**, 283–296 (2015)
- Pavlyshenko, B.: Genetic optimization of keywords subset in the classification analysis of texts authorship. *J. Quant. Linguist.* **21**(4), 341–349 (2014)
- Pemantle, R.: A survey of random processes with reinforcement *. *Prob. Surv.* **4**, 1–79 (2007)
- Pietramala, A., Policcchio, V.L., Rullo, P., Sidhu, I.: A genetic algorithm for text classification rule induction. *Lect. Notes Comput. Sci.* **5212**, 188–203 (2008)
- Pudaruth, S., Soyjaudah, K.M.S., Gunpath, R.P.: Categorisation of supreme court cases using multiple horizontal thesauri. *Intell. Syst. Technol. Appl.* **2**, 355–368 (2016)
- Read, J., Martino, L., Luengo, D.: Efficient Monte Carlo methods for multi-dimensional learning with classifier chains. *Pattern Recogn.* **47**, 1535–1546 (2014)
- Rogers, B.C.: Using genetic algorithms for feature set selection in text mining. Thesis (MSc), Miami University, Oxford, Ohio (2013)
- Roy, N., Mccallum, A.: Toward optimal active learning through sampling estimation of error reduction. In: *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 441–448 (2001)
- Sahin, I.E.: Online text categorization using genetic algorithms. Bilkent University, Turkey, Technical report, BU-CE-0704 (2007)
- Samad, S.A.: Random walk oversampling technique for minority class classification. Thesis (MSc), Tampere University of Technology (2013)
- Smith, R., Hussain, M.S.: Genetic algorithm sequential Monte Carlo methods for stochastic volatility and parameter estimation. In: *Proceedings of the World Congress on Engineering (WCE 2012)*, London, UK, 4–6 July 2012, vol. 1 (2012)
- Song, W., Park, S.C.: Genetic algorithm for text clustering based on latent semantic indexing. *Comput. Math Appl.* **57**, 1901–1907 (2009)
- ter Braak, C.J.F.: A Markov Chain Monte Carlo version of the genetic algorithm differential evolution: easy Bayesian computing for real parameter spaces. *Stat. Comput.* **16**(3), 239–249 (2006)
- Thomas, J.D., Sycara, K.: Integrating genetic algorithms and text learning for financial prediction. In: *Proceedings of the Genetic and Evolutionary Computing Conference (GECCO)*, Las Vegas, Nevada, pp. 72–75
- Waad, B., Mufti, G.B, Liman, M.: A new feature selection technique applied to credit scoring data using a ranked aggregation approach based on: optimisation, genetic algorithm and similarity. In: Osei-Bryson, K., Barclay, C. (eds.) *Knowledge Discovery Process And Methods To Enhance Organisational Performance*, pp. 347–376. CRC Press, Boca Raton (2014)
- Wang, R., Youssef, A.M., Elhakeem, A.K.: On some feature selection strategies for spam filter design. In: *Proceedings of the IEEE Canadian Conference on Electrical and Computer Engineering (CCECE 2006)*, Ottawa, Canada, 7–10 May 2006, pp. 2155–2158 (2006)
- Winands, M.H.M., Bjornsson, Y., Saito, J.T.: Monte Carlo tree search solver. In: *Proceedings of the 6th International Conference on Computers and Games*, pp. 25–36 (2008)
- WordNet: a lexical database for English. Princeton University (2017). <https://wordnet.princeton.edu/wordnet/>. Accessed 31 Jan 2017
- Wu, J., Zheng, C., Chien, C.C., Zheng, L.: A comparative study of Monte Carlo simple genetic algorithm and noisy genetic algorithm for cost-effective sampling network design under uncertainty. *Adv. Water Resour.* **29**, 899–911 (2006)
- Xiao, X.: Advanced Monte Carlo techniques: an approach for foreign exchange derivative pricing. Thesis (PhD), University of Manchester, UK (2007)

- Yang, C., Li, Y., Zhang, C., Hu, Y.: A fast KNN algorithm based on simulated annealing. In: Proceedings of the International Conference on Data Mining, Las Vegas, Nevada, 25–28 June 2007, pp. 46–51 (2007)
- Zhong, M., Shen, K., Seiferas, J.: The convergence-guaranteed random walk and its application in peer-to-peer networks. *IEEE Trans. Comput.* **57**(5), 619–633 (2008)
- Zhou, Y.: A random-walk based privacy-preserving access control for online social networks. *Int. J. Adv. Comput. Sci. Appl.* **7**(2), 74–79 (2016)
- Zhu, F., Li, H., Yao, N., Zhu, H.: Text feature selection applied by improved SAA*. *J. Comput. Inf. Syst.* **11**(17), 6419–6427 (2015)
- Zhu, H., Chen S., Pu, C., Liu, Y., Eguchi, K., Zhang, S.: Paralleling genetic annealing algorithm with OpenMP. In: Proceedings of the 2nd IEEE International Conference on Intelligent Networks and Intelligent Systems (ICINIS 2009), Tianjin, China, 1–3 November 2009

An Adaptive Soft Set Based Diagnostic Risk Prediction System

Terry Jacob Mathew^{1(✉)}, Elizabeth Sherly², and José Carlos R. Alcantud³

¹ School of Computer Sciences, Mahatma Gandhi University, Kottayam, India
terryjacobin@gmail.com

² IIITM-K, Technopark, Thiruvananthapuram, India
sherly@iiitmk.ac.in

³ BORDA Research Unit and Multidisciplinary Institute of Enterprise (IME),
University of Salamanca, 37007 Salamanca, Spain
jcr@usal.es
<http://diarium.usal.es/jcr>

Abstract. Recently, risk based prediction models in medical diagnostic systems gain wider significance in deciding most appropriate diagnostic treatments and for clinical usage. Prostate cancer is a disease which is difficult to diagnose and there are number of failure cases reported. Therefore, an effective and aggressive selection of multiple factors influence on the disease is required. In this paper, an adaptive soft set based diagnostic risk prediction system is presented with the implementation on prostate cancer. The system receives input parameters related to the disease and gives out the risk percentage of the patient. Soft sets are generated with the input parameters by fuzzification followed by rule generation. The risk percentage of the rules are individually calculated for Precision, Recall and F-Measure, that conclude on the best risk percentage based on the maximum area under the curve (AUC) in each case. This ensures to select the most influential risk parameters in treating the disease. Specificity and sensitivity of the test system yield 75.00% and 45.45% respectively.

Keywords: Soft sets · Fuzzy set · Prostate cancer · Decision making

1 Introduction

The presence of intelligent systems in the field of medical sciences have been undergoing phenomenal growth for the last two decades. Earlier, expert systems have significantly influenced the way a doctor deals with and diagnose a patient. Some notable examples are MYCIN [33], INTERNIST [25], etc. Further more, the advances in information technology embraces the digitization of the medical records and the development of other technology related medical applications in an accelerated pace.

The use of applications involving artificial intelligence and machine learning can successfully assist physicians with distinctive diagnosis of diseases, treatment opinions and recommendations, radio diagnosis on images etc. Data mining in health care also provide “real time” diagnostic as well as effective medicine recommendations on the basis of a training data set. But, driven by the issues in existing risk prediction systems, an adaptive method to improve the scope and accuracy of the prediction system is presented.

The organization of this paper is as follows. Section 2 discusses the related works. Section 3 gives an overview of the adaptive soft set based risk prediction system along with the preliminaries in Subsect. 3.1. The proposed algorithm is given in Subsect. 3.2. The implementation details of the proposed algorithm for prostate cancer detection is given in Sect. 4 along with the detailed description of each step. Section 5 discusses the results and implementation details of the algorithm. We conclude in Sect. 6.

2 Related Work

Fuzzy sets and fuzzy logic were introduced by Zadeh [38] to handle problems with uncertainty, and since then, they have contributed to a paradigm shift in the way we deal with imprecise problems and their solutions. Fuzzy sets are highly successful in many areas and give improved results than what the classical approach does. However, fuzzy set depends on membership function to represent impreciseness and are subjective to the user-level intervention. This often degrades the overall performance. Hence, to solve these problems, many researchers put forward solutions, like Atanassov [5] put forward the concept of intuitionistic fuzzy sets; Pawlak [29] introduced rough sets; Torra [35] put forward the hesitant fuzzy sets; which are applicable in real-world situations as in Alcantud et al. [1], etc. Nevertheless, these theories are limited due to the lack of parameterization concept associated with them for describing the problem.

In 1999, Molodtsov [26] introduced soft sets and established the fundamental results of this theory. A soft set is a collection of approximate descriptions of an object and is used as a general mathematical tool for dealing with objects which have been defined using a very loose and hence very general set of characteristics. Molodtsov further showed that soft set theory is free from parameterization inadequacy syndrome. Ali et al. [4] and Feng and Li [18] also contributed to settle the fundamental laws that govern this notion. Extensions and hybrid models that combine the soft set model with others have been defined and used for decision making e.g., in Ali [3], Das [11], Feng [16], Feng et al. [17, 19], Ma et al. [22], Peng and Yang [30] Zhan et al. [39] and [15].

Recently the use of soft set based intelligent systems have gained interest in computational intelligence by giving better solutions for compound problems. Soft set theory and fuzzy set theory have been successfully used in some medical systems, for example [9, 27, 28, 36]. The intelligent systems in medicine based on soft sets, generally depend on finding the risk percentage as a single value and then use it to grade the severity of the condition. This unilateral approach will not give a comprehensive picture of the real risk percentage of the patient.

We take up receiver operator characteristic (ROC) curve as a preferred performance evaluation tool to validate classifier performance over a range of decision thresholds [10,13]. The area under the curve (AUC), has been traditionally used in medical diagnosis since the 1970's [20]. The AUC maps the entire ROC curve into single number, that reflects the overall performance of the classifier over all thresholds.

Prostate cancer is the common cause of cancer death among men and it depends on various elements such as hereditary factors, age, ethnic background, the level of prostate specific antigen (PSA) in the blood etc. The level of PSA in the blood is a very important indicator for an initial diagnosis in patients [7]. But, as multiple factors cause the level of PSA to fluctuate, there exists uncertainty in the diagnosis. A biopsy of the prostate can give a distinctive diagnosis of cancer. But all patients will not be cancer positive after the biopsy. Also, it is always better to avoid an unnecessary biopsy as doctors and researchers have noted that biopsy of a tumour can cause spread of cancer cells leading to multiple sites of tumour at the biopsy site [14]. To help the doctor detect the patients with low risk of prostate cancer, a decisive intelligent system with more significant risk percentage prediction is needed.

2.1 Existing Methods and the Scope for a New Proposal

Soft computing is a host of methodologies which work in unison for providing flexible information processing capability for handling real life uncertain situations. It exploits the tolerance for imprecision, uncertainty, approximate reasoning and partial truth to build low-cost solutions. Apart from fuzzy logic, neural networks, and genetic algorithm methodologies, emergence of soft set based medical prediction systems are also on the rise. We take a few for evaluation.

Sanchez [31] initiated the use of fuzzy techniques to possibility distributions in natural languages and medical diagnosis. This notion was later extended with intuitionistic fuzzy sets by De et al. [12]. Slowiński [34] experimented with 122 patients treated for duodenal ulcer by applying rough sets to create a decision algorithm which could be used in the treatment of new ulcer patients. A pioneering prediction system for calculating the patient's prostate cancer risk given in [36]. Yuksel et al. [37] combined covering soft set and rough sets to produce soft covering based rough set and applied it for prostate cancer diagnosis. In another method given by Feng in [16], soft covering based rough sets are applied to a medical problem of calculating the risk of prostate cancer. Some other papers which deal with prostate cancer risk prediction are [6,21]. A recent contribution for glaucoma detection is given in [2]. In all the above mentioned papers, the risk percentage is calculated by a single metric based on their respective methods.

The main purpose of this paper is to provide a specific approach to improve the soft set based prediction system. Our adaptive prediction system model is tested with the data of 120 patients with prostate complaint from Selcuk University Meram Medicine Faculty [32]. We are interested in improving the accuracy of the predicted risk percentage by including the prudent use of metrics like Precision, Recall and F-measure. The Precision results depend only on the retrieved

result subsets of the actual data. It does not consider the total positives in the whole of the data. If we consider Recall for analysing the risk percentage, then the total patients are considered for risk calculation. But, the issue with Recall is that false positives cannot be discerned. To compensate for the drawbacks of Precision and Recall, we depend on F-measure. The traditional F-measure is the harmonic mean of Precision and Recall.

We propose to include the tradeoff between these metrics into the risk prediction process. In [36], the risk is calculated as follows. For example, if there are 13 patients satisfying say Rule 1, of which 4 are found to be prostate cancer positive, then the risk percentage for Rule 1 is calculated as $(4 \div 13) \times 100 = 30.76$. The patients are compared with compatible rules and the highest risk percentage is accorded as the risk percentage of the patients. Here, the drawback of using Precision can be explained by this assumption. Assume that out of the compatible rules generated for a patient; say Rule 1 has only 1 patient and this patient is also tested positive in biopsy, then the patient will be awarded with 100% risk percentage. This may not be always true.

As there is no qualitative and quantitative analysis on the result output, these single handed general methods of deriving the risk percentage from soft set based and other prediction techniques are not a legitimate way of calculation. Therefore, we propose to include Precision, Recall and F-measure in finding the significant associated risk. Also, rather than taking the highest observed risk percentage of a specific rule as the risk of the patient, an average of the most relevant risks are calculated. This averaging of the selected rules make our model more effective.

3 Adaptive Soft Set Based Risk Prediction System

In this section, we present the basic definitions of soft set theory [26] and fuzzy set theory [23] followed by the proposed system. These definitions and further details on soft sets and fuzzy sets can be found in [8, 24, 38]. As usual, we follow the common terminology for describing soft set and its extensions. Here U refers to an initial universe and E is the set of parameters.

3.1 Definitions: Soft Set and Fuzzy Set

Definition 1 (Molodtsov [26]). A pair (F, A) is a soft set over U when $A \subseteq E$ and $F : A \rightarrow \mathcal{P}(U)$, where $\mathcal{P}(U)$ denotes the set of all subsets of U .

Example 1. A soft set over U is regarded as a parameterized family of subsets of the universe U , the set A being the parameters. For each parameter $e \in A$, $F(e)$ is the subset of U approximated by e or the set of e -approximate elements of the soft set.

Let U be the set of five patients given by $U = \{p_1, p_2, p_3, p_4, p_5\}$ and E be the set of symptoms given by $E = \{s_1, s_2, s_3, s_4, s_5\}$.

Let $A = \{s_1, s_2, s_3\}$ be the set of symptoms, the doctor intends to use for diagnosis. Now consider that (F, A) is a mapping given by, $(F, A)(s_1) = \{p_1, p_2\}$, $(F, A)(s_2) = \{p_1, p_3\}$ and $(F, A)(s_3) = \{p_2, p_4\}$.

Then the soft set $(F, E) = \{(s_1, \{p_1, p_2\}), (s_2, \{p_1, p_3\}), (s_3, \{p_2, p_4\}), (s_4, \{\emptyset\}), (s_5, \{\emptyset\})\}$. A soft set can also be represented in the form of a two dimensional table. Table 1 is the tabular representation of the soft set (F, E) shown in Example 1.

Table 1. Tabular representation of the fuzzy set (F, A) associated with Example 1.

U/E	s_1	s_2	s_3	s_4	s_5
p_1	1	1	0	0	0
p_2	1	0	1	0	0
p_3	0	1	0	0	0
p_4	0	0	1	0	0
p_5	0	0	0	0	0

Definition 2 (Maji et al. [24]). Let (F, A) and (G, B) be two soft sets. Then the AND operation of (F, A) AND (G, B) , denoted by $(F, A) \wedge (G, B)$, is defined as $(H, A \times B)$ where $H(\alpha, \beta) = F(\alpha) \cap G(\beta)$ for each $(\alpha, \beta) \in A \times B$.

Definition 3 (Zadeh [38]). A fuzzy set X over U is a set defined by a function μ_X representing a mapping $\mu_X : U \rightarrow [0, 1]$. where, μ_X is called the membership function of X , and the value $\mu_X(u)$ is the grade of membership of $u \in U$. The value of $\mu_X(u)$ represents the degree with which u belongs to the fuzzy set X . Thus, a fuzzy set X over U can be represented as follows:

$$X = \{(\mu_X(u)/u) : u \in U, \mu_X(x) \in [0, 1]\}.$$

For a fuzzy set X in U and any real number $\alpha \in [0, 1]$, then the α -cut or cut worthy set of A , denoted by $X[\alpha]$ is the crisp set defined as $\{x \in U : \mu_X(x) \geq \alpha\}$.

3.2 Proposed Intelligent System for Prostate Cancer Diagnosis

The available data set is attributed with a set of three variables, namely prostate specific antigen (PSA), prostate volume (PV) and age of the patient. The membership function of these variables are shown in Eqs. (1) and (2). All 120 selected patients underwent biopsy and their diagnostic results are known. In the following part, we proceed to explain the step by step procedures, which make up the proposed algorithm.

In order to facilitate the representation of soft sets, we initially convert the input data into fuzzy sets. Afterwards, going by the principle of including fuzzy

sets as soft sets (cf., Molodtsov [26]), the fuzzy sets are redeployed correspondingly as relevant soft sets. Unlike the conventional soft set prediction methods, we avoided parameter reduction in view of the nature and type of data set. The decisive phase is generation of rules, which are analysed later for determining the prostate cancer risk. Each rule is awarded a risk percentage which determines the verdict of the intelligent system. The algorithm for prostate cancer detection with stepwise descriptions is given below.

AdaptiVe Algorithm for Softset based predicTion (AVAST).

- Step 1.** Fuzzyfication of data set with the selected variables namely PSA, PV and age.
- Step 2.** Transforming the fuzzy sets corresponding to input data into soft sets.
- Step 3.** Obtaining the rules relevant for the system by the application of AND operator on to the soft sets generated in the previous step.
- Step 4.** Analysis of rules based on Precision, Recall and F-measure.
- Step 5.** Plot the ROC curve with the calculated risk percentage for the above three sets.
- Step 6.** Select the metric i.e. (either Precision, Recall or F-measure), which offers the maximum AUC and proceed for actual risk prediction over the testing set.

4 Implementation of Algorithm - AVAST to Calculate Prostate Cancer Risk

The various stages of algorithm - AVAST is explained in detail below.

Explanation of Step 1. We fuzzificate the patient data with appropriate membership functions on the basis of inputs from medical literature [36]. The following linguistic variables are modelled for the attributes PSA, PV and age. The PSA variables VL, L, M, H and VH represent very low, low, middle, high and very high respectively. The PV variables S, M, B and VB represent small, medium, big and very big respectively. The age factor attributed by VY, Y, M and O represents very young, young, middle and old respectively. Trapezoidal or triangular membership functions can be selected for each variable on the basis of their interval size. The corresponding membership values are determined from Eqs. 1 and 2.

$$PSA(x) = \begin{cases} \mu_x & \text{if } 0 < x < 100 \\ 1 & \text{if } x \geq 100 \end{cases} \quad PV(y) = \begin{cases} \mu_y & \text{if } 30 < y < 120 \\ 1 & \text{if } y \geq 120 \end{cases} \quad (1)$$

$$Age(z) = \begin{cases} 0 & \text{if } z \leq 20 \\ \mu_z & \text{if } 20 < z < 65 \\ 1 & \text{if } z \geq 65 \end{cases} \quad (2)$$

Table 2. A sample input data of patients

U	Age	PSA	PV
U_7	54	5.62	28
U_9	54	17.3	45
U_{20}	59	8.36	55
U_{34}	61	18.3	62
U_{40}	62	51.74	29
U_{70}	68	140	117
U_{99}	73	47.4	87

A sample of the input data is shown in Table 2 and the parameter memberships are shown in Fig. 1.

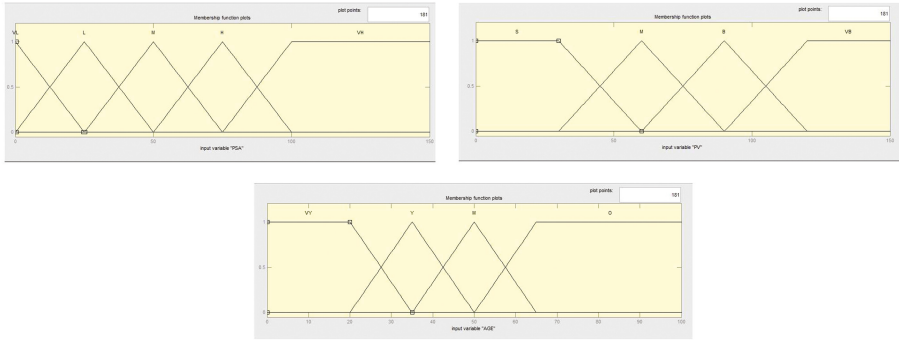


Fig. 1. The membership functions of Age, PSA and PV

Explanation of Step 2. We directly depend on Molodtsov’s method for the transformation of fuzzy sets into soft sets. The Molodtsov’s method maps soft sets on to the universe $[0, 1]$, thus making the selection of a subset of this range inevitable to conduct a practical setting of this experiment. Depending on the distribution of patient data into different levels of membership, we have different elements in different subsets for each variables. Table 3 shows the fuzzy membership values of the input factors. In this approach, for a soft set (F, A) over U , A is denoted by a set of parameters represented by $\{e_1, e_2, e_3, e_4, \dots, e_n\}$, then $F(e_j)$ is a subset of U , where, $F(e_j) = \{U_i \mid \mu(U_i) \geq e_j; \forall j = 1 \text{ to } n \text{ and } i = 1 \text{ to } m\}$.

Hence, the newly formed soft sets will have subsets of elements from the universal set U . As an example from the data set, the soft set,

Table 3. The fuzzy membership values of factors shown in Table 2

U	Age	PSA	PV
U_7	0.73 M, 0.26 O	0.77 VL, 0.22 L	1 S
U_9	0.73 M, 0.26 O	0.30 VL, 0.69 L	0.5 S, 0.5 M
U_{20}	0.4 M, 0.6 O	0.66 VL, 0.33 L	0.16 S, 0.83 M
U_{34}	0.26 M, 0.73 O	0.26 VL, 0.73 L	0.93 M, 0.06 B
U_{40}	0.2 M, 0.8 O	0.93 M, 0.06 H	1 S
U_{70}	1 O	1 VH	0.1 B, 0.9 VB
U_{99}	1 O	0.10 L, 0.89 M	0.1 M, 0.9 B

$F : A_{Age(M)} \longrightarrow \mathcal{P}(U)$ is associated with a parameter set,

$A_{Age(M)} = \{0.06, 0.28, 0.5, 0.71, 0.93\}$, and the corresponding soft sets obtained are as:

$$F(.06) = \{U_{42}, U_{43}, U_{44}, U_{45}, U_{46}, U_{41}, U_{35}, U_{37}, U_{40}, U_{30}, U_{31}, U_{32}, U_{33}, U_{34}, U_{25}, U_{27}, U_{28}, U_{29}, U_{19}, U_{20}, U_{21}, U_{22}, U_{23}, U_{24}, U_{15}, U_{16}, U_{17}, U_{18}, U_1, U_{13}, U_{10}, U_8, U_9, U_4, U_5, U_2, U_3\}$$

$$F(.28) = \{U_{25}, U_{27}, U_{28}, U_{29}, U_{19}, U_{20}, U_{21}, U_{22}, U_{23}, U_{24}, U_{15}, U_{16}, U_{17}, U_{18}, U_1, U_{13}, U_{10}, U_8, U_9, U_4, U_5, U_2, U_3\}$$

$$F(.5) = \{U_1, U_{13}, U_{10}, U_8, U_9, U_4, U_5, U_2, U_3\},$$

$$F(.71) = \{U_8, U_9, U_4, U_5, U_2, U_3\}, \text{ and}$$

$$F(.93) = \{U_2, U_3\}$$

Explanation of Step 3. The combination of soft sets obtained in Step 2 by AND'ing operation gives all possible rules. By this means, a total of 1760 rules are generated, which are checked for compatibility with the patients. An obtained sample rule is given below.

For example: $AGE(M)(.5) \wedge PSA(VL)(.05) \wedge PV(M)(.5) = \{u_{10}, u_{14}\}$.

Explanation of Step 4. The output obtained from Step 3 is processed further to associate each rule with a risk of prostate cancer as follows.

The rules obtained above will have patients from the training data and the Precision, Recall and F-measure based risk percentage is calculated for each rule. The calculated risk percentage of patients are then separately considered and averaged individually. Thus, corresponding to each test data, we now have separate risk percentage for Precision, Recall and F-measure.

For a sample training data, the rules obtained are as:

$$\begin{aligned} R_1 &= \{U_{10}, U_{14}\} \\ R_2 &= \{U_{13}, U_{17}, U_{81}\} \\ R_3 &= \{U_{21}, U_{22}, U_{54}, U_{67}, U_{98}, U_{107}\} \\ R_4 &= \{U_3, U_8, U_{21}, U_{22}, U_{54}, U_{67}, U_{98}, U_{99}, U_{107}\} \text{ and} \\ R_5 &= \{U_{40}, U_{92}, U_{116}\} \end{aligned}$$

The biopsy results for the patients are known from the labelled data set and the calculated values of Precision, Recall and F-measure (F1) for the above rules are calculated as per Eq. 3 and are shown in the Table 4. Here TP indicates true positive, FN is false negative and FP is false positive.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, F1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (3)$$

Some of the validation results of rules' risk percentage based on Precision, Recall and F-measure are shown in Table 5, where test samples with "*" indicate the outliers.

Table 4. The risk percentage of rules on the basis of Precision, Recall and F-Measure

Rules	Precision	Recall	F-Measure
R_1	0	0	0
R_2	0	0	0
R_3	0.5	0.06	0.11
R_4	0.56	0.1	0.18
R_5	1	0.06	0.12

Table 5. The risk percentage of a set of patients by Precision, Recall and F-measure

Test data	Risk percentage while considering			Ground truth
	Precision	Recall	F-Measure	
U_1	38.29	10.69	7.92	0
U_{10}	38.36	13.35	8.81	0
U_{31}	43.27	17.68	11.37	0
U_{43}	47.91	28.5	16.83	0
U_{60}	54.08	30.17	18.66	1
U_{77*}	41.52	5.68	4.4	1
U_{83}	55.35	22.35	14.49	1
U_{86}	54.35	28.57	17.84	1
U_{95*}	54.51	25.08	16.31	0
U_{113*}	55.09	21.55	14.47	0

The next step will determine, whether we will consider Precision, Recall or F-measure, for calculating the actual risk percentage for each patient.

Explanation of Step 5. Plot the ROC curve for the risk percentage based on Precision, Recall and F-measure and select the risk percentage for the test patients corresponding to the maximum AUC. The AUC maps the entire ROC curve into single value that portrays the overall performance of the classifier over all thresholds. The false positive rate (FPR) and true positive rate (TPR) evaluate performance for a specific threshold. The FPR and TPR can also be combined to form an overall mis-classification rate, which is known as true error. By getting the ROC, AUC, FPR and TPR of this system, we obtain the complete knowledge about the performance of the system. Table 6 shows the AUC generated for some test groups.

Table 6. The AUC of some samples used in generating the ROC curve

Test data sets	Area Under the Curve (AUC)		
	Precision	Recall	F-Measure
Dataset 1	40.15	56.43	56.43
Dataset 2	60.07	58.68	59.72
Dataset 3	50.69	68.75	53.13
Dataset 4	45.49	49.31	48.96
Dataset 5	63.54	46.53	44.44

By employing a five-fold cross validation, the patient data is randomly divided into training and testing sets. By this approach [13], the original data is randomly partitioned into five equal sized sub-samples. Of the five sub-samples, a single sub-sample is retained as the validation data for testing the model, and the remaining four sub-samples are used as training data. The cross-validation process is then repeated five times (the folds). The five results from the folds can then be averaged to produce a single estimation.

Explanation of Step 6. Finally, the choice of the appropriate metric is done from Precision, Recall and F-measure on the basis of the maximum AUC generated over the validation data.

5 Results and Discussion

After five fold cross validation, the ROC plots corresponding to Precision, Recall and F-Measure for validation data are shown in Fig. 2. The selected metric can be either Precision, Recall or F-Measure based on the AUC obtained for each case. In this investigated case, as seen in Fig. 2, Precision based ROC curve has the highest AUC than Recall and F-Measure. Hence Precision will be selected for calculating the rule risk percentage.

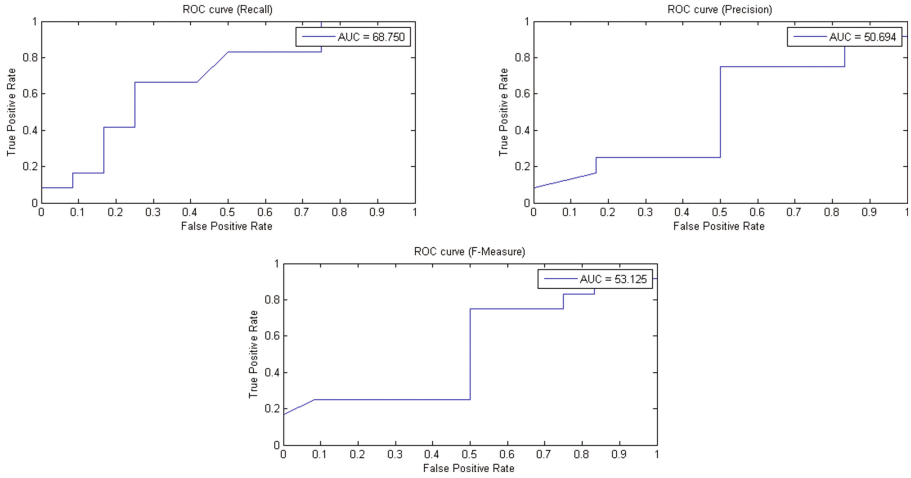


Fig. 2. The ROC curve for Precision, Recall and F-measure

The proposed algorithm model was implemented by means of scientific computation platform R2013a Matlab. In this investigation, we have divided the total data into testing data, training data and validation data. One set is for training, one for testing and the other two for validation. By the application of AVAST algorithm on the validation data, we could select the best from the Precision based, Recall based and F-Measure based methods on the basis of maximum AUC.

As sensitivity is significant for this specific case of prostate cancer prediction, we have to select a threshold value which gives high sensitivity over the validation process. Concurrently the false positive rate (FPR) should be minimal. So a particular risk percentage value is selected as the threshold when the corresponding TPR is at least 80% and with minimum FPR. This threshold is then applied for the final test data. Figure 3 shows the ROC curve for the test system. It should be noted that we can change the threshold selection parameters for

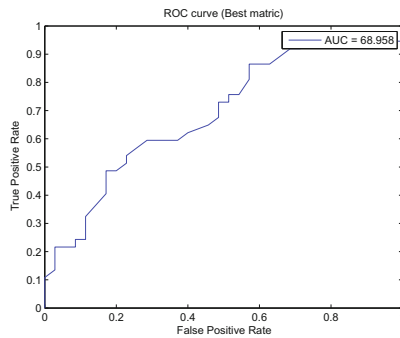


Fig. 3. The ROC curve for the test data

the system on the basis of data set and other requirements. The sensitivity and specificity of the test system stood at 75.00% and 45.45 % respectively.

6 Conclusions

In this work, it is shown that the soft set approach for finding the risk percentage of prostate cancer patients can be significantly improved with the inclusion of Precision, Recall and F-measure based rule analysis. The proposed method exhibits an adaptive nature as the best performing metric is chosen on the basis of the validation data set performance. We depend on these statistical measures to optimize the risk percentage calculation. This general notion can be applied to all methods which follow a unidirectional approach in defining the output risk percentage. The results confirm that the inclusion of this adaptive approach to existing methodologies show better results as shown in Sect. 4.

As future enhancements, we propose to extend our approach with other existing algorithms with more relevant parameters for reducing ambiguity. Also a weighted approach for rules and parameters can be employed to see if it leads to further improvements. The medical applications using the concepts of soft sets opens up lot of room for exploration and innovation. A quick and automated method of prostate cancer diagnosis based on soft sets is addressed in this contribution. The proposed model helps the doctor to discern the patients for the biopsy procedure to detect prostate cancer.

References

1. Alcantud, J.C.R., de Andres Calle, R., Torrecillas, M.J.M.: Hesitant fuzzy worth: an innovative ranking methodology for hesitant fuzzy subsets. *Appl. Soft Comput.* **38**, 232–243 (2016)
2. Alcantud, J.C.R., Santos-García, G., Hernández-Galilea, E.: Glaucoma diagnosis: a soft set based decision making procedure. In: *Conference of the Spanish Association for Artificial Intelligence*, pp. 49–60. Springer (2015)
3. Ali, M.: A note on soft sets, rough soft sets and fuzzy soft sets. *Appl. Soft Comput.* **11**, 3329–3332 (2011)
4. Ali, M.I., Feng, F., Liu, X., Min, W.K., Shabir, M.: On some new operations in soft set theory. *Comput. Math. Appl.* **57**(9), 1547–1553 (2009)
5. Atanassov, K.: Intuitionistic fuzzy sets. *Fuzzy Sets Syst.* **20**, 87–96 (1986)
6. Benecchi, L.: Neuro-fuzzy system for prostate cancer diagnosis. *Urology* **68**(2), 357–361 (2006)
7. Catalona, W.J., Partin, A.W., Slawin, K.M., Brawer, M.K., Flanigan, R.C., Patel, A., Richie, J.P., Walsh, P.C., Scardino, P.T., Lange, P.H., et al.: Use of the percentage of free prostate-specific antigen to enhance differentiation of prostate cancer from benign prostatic disease: a prospective multicenter clinical trial. *Jama* **279**(19), 1542–1547 (1998)
8. Çağman, N., Enginoğlu, S.: Soft set theory and uni-int decision making. *Eur. J. Oper. Res.* **207**(2), 848–855 (2010)
9. Çelik, Y., Yamak, S.: Fuzzy soft set theory applied to medical diagnosis using fuzzy arithmetic operations. *J. Inequalities Appl.* **2013**(1), 82 (2013)

10. Cohn, T.E.: Receiver operating characteristic analysis of photoreceptor sensitivity. *IEEE Trans. Syst. Man Cybern.* **5**, 873–881 (1983)
11. Das, A.K.: Weighted fuzzy soft multiset and decision-making. *Int. J. Mach. Learn. Cybern.* 1–8 (2016). Springer
12. De, S.K., Biswas, R., Roy, A.R.: An application of intuitionistic fuzzy sets in medical diagnosis. *Fuzzy Sets Syst.* **117**(2), 209–213 (2001)
13. D’Errico, G.E.: Receiver operating characteristic: a tool for cell confluence estimation. In: 2015 IEEE International Symposium on Medical Measurements and Applications (MeMeA), pp. 576–579. IEEE (2015)
14. Eriksson, M., Reichardt, P., Hall, K.S., Schütte, J., Cameron, S., Hohenberger, P., Bauer, S., Leinonen, M., Reichardt, A., Davis, M.R., et al.: Needle biopsy through the abdominal wall for the diagnosis of gastrointestinal stromal tumour-does it increase the risk for tumour cell seeding and recurrence? *Eur. J. Cancer* **59**, 128–133 (2016)
15. Fatimah, F., Rosadi, D., Hakim, R.F., Alcantud, J.C.R.: Probabilistic soft sets and dual probabilistic soft sets in decision-making. In: *Neural Computing and Applications*, pp. 1–11 (2017)
16. Feng, F.: Soft rough sets applied to multicriteria group decision making. *Ann. Fuzzy Math. Inform.* **2**(1), 69–80 (2011)
17. Feng, F., Li, C., Davvaz, B., Ali, M.: Soft sets combined with fuzzy sets and rough sets: a tentative approach. *Soft Comput.* **14**(9), 899–911 (2010)
18. Feng, F., Li, Y.: Soft subsets and soft product operations. *Inf. Sci.* **232**, 44–57 (2013)
19. Feng, F., Liu, X., Leoreanu-Fotea, V., Jun, Y.B.: Soft sets and soft rough sets. *Inf. Sci.* **181**(6), 1125–1137 (2011)
20. Huang, J., Ling, C.X.: Using AUC and accuracy in evaluating learning algorithms. *IEEE Trans. Knowl. Data Eng.* **17**(3), 299–310 (2005)
21. Keles, A., Hasiloglu, A.S., Keles, A., Aksoy, Y.: Neuro-fuzzy classification of prostate cancer using NEFCLASS-J. *Comput. Biol. Med.* **37**(11), 1617–1628 (2007)
22. Ma, X., Liu, Q., Zhan, J.: A survey of decision making methods based on certain hybrid soft set models. *Artif. Intell. Rev.* **47**(4), 507–530 (2017)
23. Maji, P., Biswas, R., Roy, A.: Fuzzy soft sets. *J. Fuzzy Math.* **9**, 589–602 (2001)
24. Maji, P., Biswas, R., Roy, A.: Soft set theory. *Comput. Math. Appl.* **45**, 555–562 (2003)
25. Miller, R.A., Pople Jr., H.E., Myers, J.D.: Internist-I, an experimental computer-based diagnostic consultant for general internal medicine. *New Engl. J. Med.* **307**(8), 468–476 (1982)
26. Molodtsov, D.: Soft set theory - first results. *Comput. Math. Appl.* **37**, 19–31 (1999)
27. Oniśko, A., Druzdzel, M.J.: Impact of precision of bayesian network parameters on accuracy of medical diagnostic systems. *Artif. Intell. Med.* **57**(3), 197–206 (2013)
28. Park, K.S., Chae, Y.M., Park, M.: Developing a knowledge-based system to automate the diagnosis of allergic rhinitis. *Biomed. Fuzzy Hum. Sci. Official J. Biomed. Fuzzy Syst. Assoc.* **2**(1), 9–18 (1996)
29. Pawlak, Z.: Rough sets. *Int. J. Comput. Inf. Sci.* **11**(5), 341–356 (1982)
30. Peng, X., Yang, Y.: Algorithms for interval-valued fuzzy soft sets in stochastic multi-criteria decision making based on regret theory and prospect theory with combined weight. *Appl. Soft Comput.* **54**, 415–430 (2017)
31. Sanchez, E.: Inverses of fuzzy relations. Application to possibility distributions and medical diagnosis. *Fuzzy Sets Syst.* **2**(1), 75–86 (1979)

32. Saritas, I., Allahverdi, N., Sert, I.U.: A fuzzy approach for determination of prostate cancer. *Int. J. Intell. Syst. Appl. Eng.* **1**(1), 1–7 (2013)
33. Shortliffe, E.: *Computer-Based Medical Consultations: MYCIN*, vol. 2. Elsevier, New York (2012)
34. Slowinski, K.: Rough classification of HSV patients. In: *Intelligent Decision Support-Handbook of Applications and Advances of the Rough Sets Theory*, pp. 77–94 (1992)
35. Torra, V.: Hesitant fuzzy sets. *Int. J. Intell. Syst.* **25**(6), 529–539 (2010)
36. Yuksel, S., Dizman, T., Yildizdan, G., Sert, U.: Application of soft sets to diagnose the prostate cancer risk. *J. Inequalities Appl.* **2013**(1), 229 (2013)
37. Yüksel, Ş., Tozlu, N., Dizman, T.H.: An application of multicriteria group decision making by soft covering based rough sets. *Filomat* **29**(1), 209–219 (2015)
38. Zadeh, L.: Fuzzy sets. *Inf. Control* **8**, 338–353 (1965)
39. Zhan, J., Liu, Q., Herawan, T.: A novel soft rough set: soft rough hemirings and corresponding multicriteria group decision making. *Appl. Soft Comput.* **54**, 393–402 (2017)

Weighted Bipartite Graph Model for Recommender System Using Entropy Based Similarity Measure

Punam Bedi, Anjali Gautam^(✉), Saumya Bansal, and Deepika Bhatia

Department of Computer Science, University of Delhi, Delhi, India
punambedi@ieee.org, agautam@cs.du.ac.in,
saumya.mcs.du.2015@gmail.com,
deepika.mcs.du.2015@gmail.com

Abstract. Collaborative filtering technique is widely adopted by researchers to generate quality recommendations. Constant efforts are being made by the researchers to generate quality recommendations thus satisfying and retaining the user. This work is an effort to generate quality recommendations by proposing a collaborative filtering approach. The proposed work models the sparse rating data as a weighted bipartite graph which represents data flexibly and exploits the graph properties to generate recommendations. In the proposed work user similarity is formulated as measure of entropy and cosine similarity which takes into account the relative difference between the ratings. Performance of the proposed approach is compared with the traditional collaborative filtering technique using Precision, Recall and F-Measure. Experiments were conducted on public and private datasets namely MovieLens and News dataset respectively. Results indicate that the performance of the proposed approach outperforms the traditional collaborative filtering approach.

Keywords: Collaborative filtering · Weighted bipartite graph · Information entropy

1 Introduction

Recommender System (RS) is an efficient software system that helps the user to navigate swiftly in the era of data explosion. In order to generate recommendations, RS tends to find user preferences from the history or feedback from the target user and identifies similar set of users based on preferences of the target user. This technique of recommendation is termed as Collaborative Filtering (CF) which is the basis of our work.

The technique of CF is posed with a limitation of data sparsity (Jannach 2010). The problem of data sparsity arises due to lack of inadequate information about users and items. Precisely, when the users are not willing to rate the items; it becomes difficult to find the similar set of users thus making the recommendation tasks much more difficult. Consider a situation where a user rarely visits RS, as a result of which he/she rates only few items from an entire huge list of items; or a user who frequently visits the system, get the recommendations but does not really like to rate the items. The above scenarios are a perfect case of data sparsity as the system knows less about the likes and

preferences of the user. Generating quality recommendations in such scenarios becomes very difficult.

Aim of the proposed work is to improve the quality of recommendations generated from the sparse rating matrix by representing the sparse rating matrix using a weighted bipartite graph and thereafter exploiting the bipartite graph properties to generate recommendations. A weighted bipartite graph is created using the sparse matrix where the two sets of nodes are user nodes and item nodes respectively. Bipartite graphs are capable of representing data flexibly by showing the presence of an edge between two nodes (user, item) which corresponds that an item has been rated by the user and the weight of the edge represents the rating an item has received by the user. To make the users collaborate in CF technique, user similarity is computed based on hybrid similarity metric which is a sum of cosine similarity measure (Jannach 2010) and information entropy (Piao et al. 2009).

The paper is structured into distinct sections which are as follows: Sect. 2 discusses the relevant related work. The proposed approach is detailed in Sect. 3. Section 4 discusses the experiments and results. Section 5 concludes the paper.

2 Related Work

Researchers have widely adopted the technique of CF to generate recommendations for the target user. Constant efforts are being made to generate quality recommendation by overcoming the limitations of CF. This section discusses the relevant related work.

2.1 Graph-Based Recommender System

Many researchers have incorporated the use of graphs and exploited graph properties in order to generate recommendations. These graphs can be of varied types ranging from normal graphs (Lee and Lee 2015; Huang et al. 2002) to bipartite (Chen et al. 2011, 2013; Huang et al. 2004; Sawant 2013; Lopes et al. 2016) and even tripartite graphs (Shams and Haratizadeh 2017). Chen et al. (2013) proposed a graph based approach of recommendation. Recommendation for the target user is projected as a problem of dynamic resource allocation combined with computing item similarity. Each entry in the resource allocation matrix represents the resources, one item would like to distribute to the other item. Finally resource allocation for the target user can be determined by combining the similarity measure along with initial resource allocation of the target user. Huang et al. (2004) and Chen et al. (2011) created a bipartite graph and used association retrieval by exploiting the path length property to generate recommendations. Sawant (2013) in his work used the weighted bipartite graph projection that defined a new similarity function based on the network properties of the dataset which was then used to generate recommendations. A 2-step walk from one user to other was considered for computing user similarity. Lee and Lee (2015) used graph based approach to generate novel recommendations by considering only the positively rated items. Entropy is used by the authors to separate popular and novel items from the entire set of items. Not only bipartite graphs have been used for recommendation, Shams and Haratizadeh (2017) studied a way where tri-partite graph structure was used

to capture preference data which was then explored to capture the different kinds of relations existing in a ranking preference dataset in order to generate recommendations. Lopes et al. (2016) proposed a computationally efficient graph based collaborative filtering approach based on short path enumeration property for binary ratings. A hybrid recommendation approach based on weighted bipartite graph and item based CF was proposed by Hu et al. (2016) to solve the problem of data sparsity. Use of weighted bipartite graph facilitated resource allocation evenly.

2.2 Entropy Based Similarity Metric in RS

To improve the accuracy of recommendations researchers have used the concept of entropy based similarity metrics to compute either similar items (Piao et al. 2009) or similar users (Wang et al. 2015) or both. Entropy based prediction coefficient is also proposed in Chandrashekhar and Bhasker (2011).

Entropy based item-item similarity was proposed by Piao et al. (2009). To generate item based similarity, joint entropy for the set of items was considered. For large number of items, it is difficult to compute entropy of each item and joint entropy between items as it is computationally expensive. On the other hand, Wang et al. (2015) proposed a recommendation technique which employed entropy based user similarity measure and Manhattan distance to generate recommendations. Chandrashekhar and Bhasker (2011) proposed a memory based CF technique based on the idea of selective predictability and made use of entropy to estimate it which outperformed the traditional CF technique. Another approach was proposed by Mehta et al. (2011) which dealt with the scalability issue which occurred due to increase in information domain on the internet. The proposed approach tried to reduce the scalability issue by computing user similarity using entropy based similarity measure. By using the entropy between all the users, the approach filtered trustworthy users which were used to generate recommendations for the target user. Entropy between users was based on the score rating difference between the target user and other user for n different pages.

RS based on graphs has been proposed by a number of researchers as is mentioned in this section. It was found from the literature review that varieties of graphs are used to generate quality recommendations from the sparse rating matrix using the technique of CF. The drawbacks of these approaches lies in the interpretation of similarity measure to generate predictions (Huang et al. 2002). This paper proposes a CF approach based on weighted bipartite graph. The proposed approach exploits the properties of a weighted bipartite graph to generate recommendations for the target user. To keep the essence of similarity in the CF technique, the proposed approach makes use of a hybrid user similarity metric which is formulated based on the concept of entropy in addition to the standard cosine similarity measure.

3 Proposed Work

This paper proposes a CF approach based on weighted bipartite graph. The proposed approach exploits the properties of a weighted bipartite graph to generate recommendations for the target user. To keep the essence of collaboration in the recommendation technique, the proposed approach makes use of a hybrid user similarity metric which is formulated based on the concept of entropy in addition to the standard cosine similarity measure. This section details the proposed approach.

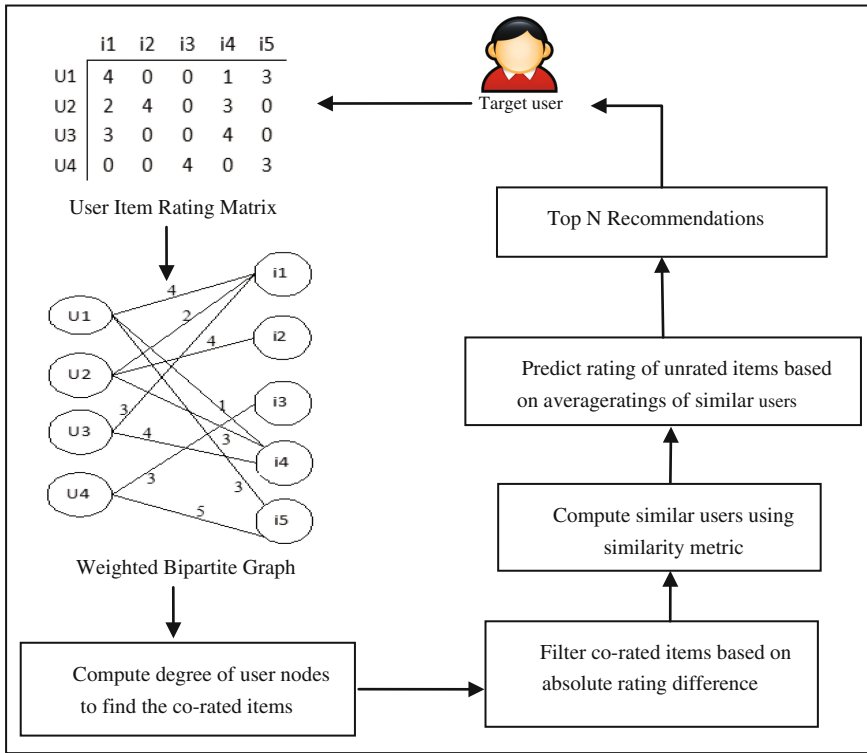


Fig. 1. Proposed framework.

3.1 Proposed Approach

In CF, user-item interaction is captured in the form of sparse user-item rating matrix R where each entry r_{ij} represents the rating given by the user i to item j . An efficient way to generate recommendations from this sparse matrix is to represent matrix in the form of weighted bipartite graph. Bipartite graph represents data flexibly, reducing the effect of sparsity in the data. The main objective of the proposed work is to generate quality recommendations using memory based collaborative filtering approach as opposed to model based CF. Figure 1 depicts the framework of the proposed approach for a target user. Each step of the proposed approach is detailed below.

Step 1: Build a Bipartite Graph G

Let there be N users and M items. Using the $N * M$ rating matrix we establish the weighted bipartite graph $G = (N, M, E)$ with $N + M$ nodes and E being the edge set to specify the preferences of the users for a particular item. There exists an edge from a node in the user set to a node in the item set if the user has rated the corresponding item with a weight equal to the rating given by user to the item. In Fig. 1, the equivalent bipartite graph corresponding to the rating matrix consists of 4 user nodes and 5 item nodes and the set of weighted edges.

Step 2: Determining commonly rated items using degree of each node

Commonly rated items between the target user U_i and $N - 1$ users are determined by exploiting the degree property of the nodes in the graph G . Let $\text{degree}(U_i)$ and $\text{degree}(U_j)$ denotes the degree of the target user U_i and $U_j \in N - \{U_i\}$ respectively in G . Co - rated items $_{(U_i, U_j)}$ is the count of maximum number of commonly rated items between user U_i and U_j represented by Eq. (1).

$$\text{Co - rated items}_{(U_i, U_j)} = \min(\text{degree}(U_i), \text{degree}(U_j)) \quad (1)$$

For each item rated by the user node (U_i/U_j) with minimum degree, the system determines whether there exists an edge between the other user node (U_j/U_i) for the corresponding item in G . By keeping track of all the edges that exists from the user node with more degree to all the items rated by the user node with minimum degree the system determines the set of commonly rated items.

Step 3: Filter co-rated items based on absolute rating difference

In the previous step the system determines the commonly rated items between the two users which becomes the preliminary step to determine the similar set of users to the target user in the proposed approach. Users are said to be similar if they rate some common set of items. To improve the user-user similarity and hence improving the efficiency of the system it is important to take the rating difference of the commonly rated item rather than just considering set of common items. This difference in the rating will help the system to identify the level of similarity between the two users. For each of the co-rated items between user U_i and U_j determined in Step 2, compute absolute rating difference given by $|r_{i,k} - r_{j,k}|$ where $r_{i,k}$ denotes rating of user U_i to item k , $r_{j,k}$ denotes rating of user U_j to item k and $k \in [1 \dots \text{Co - rated items}_{(U_i, U_j)}]$. For filtering the number of items among the commonly rated items between two users, we assign weights to the count of the absolute rating differences giving the maximum weight to minimum absolute rating difference because the items with less absolute rating difference should have more influence on similarity computation and vice-versa. Equation (2) formulates the criteria for filtering the items from the Co - rated items $_{(U_i, U_j)}$.

$$p(U_i, U_j) = 1 \times \text{count}[0] + 0.8 \times \text{count}[1] + 0.6 \times \text{count}[2] + 0.4 \times \text{count}[3] + 0.2 \times \text{count}[4], \quad (2)$$

where $count[i]$ denotes the number of items with absolute rating difference i and 1, 0.8, 0.6, 0.4 and 0.2 are the assigned weights. $p(U_i, U_j)$ denotes the weighted sum of the absolute rating difference.

$$t(U_i, U_j) = \frac{P(U_i, U_j)}{\text{Co - rated items}_{(U_i, U_j)}} \times 100 \tag{3}$$

$t(U_i, U_j)$ be the percentage of items with minimum absolute difference among Co - rated items $_{(U_i, U_j)}$ to be considered for the user similarity computation based on sorted values of absolute rating difference.

Step 4: Compute similar users to the target user

After filtering the co-rated items, the system determines the set of similar users to the target user. To keep the essence of CF intact while using the weighted bipartite graph this step computes the similar users to the target user using a hybrid similarity metric (Wang et al. 2015). The similarity metric is expressed as a sum of entropy measure and cosine similarity measure. In information retrieval, information entropy $H(Y)$ is a measure of uncertainty associated in a random variable Y with values ranging from $\{y_1, y_2, \dots, y_n\}$ and is expressed as in Eq. (4)

$$H(Y) = - \sum_{i=1}^n P(y_i) \log_2 P(y_i) \tag{4}$$

where $P(y_i)$ is the probability function of the random variable.

In our problem entropy is used to find similar users to the target user based on the item set I filtered in the above step. In our case, the random variable y_i is the item $k \in [1..I]$. The probability function is calculated using the weights of the edges in G . The probability function is based on the relative difference of deviation of the ratings in addition to absolute difference. For users U_i and U_j , deviation between them for item k is computed as

$$de_k(U_i, U_j) = |(r_{i,k} - \bar{r}_i) - (r_{j,k} - \bar{r}_j)| \tag{5}$$

where \bar{r}_i, \bar{r}_j denotes average rating of user U_i and user U_j respectively; $r_{i,k}$ and $r_{j,k}$ denotes rating of item k by user U_i , and user U_j respectively. Item $k \in [1..I]$ where I denote the number of co-rated items filtered in step 3. Deviation across all the co-rated items for the user U_i and U_j is represented as in Eq. (6)

$$\sum_{k=1}^I de_k(U_i, U_j) \tag{6}$$

The probability function of a random variable is defined as

$$P_k = \frac{de_k(U_i, U_j)}{\sum_{k=1}^I de_k(U_i, U_j)} \tag{7}$$

The information entropy (H) for our problem is defined as (8)

$$H = - \sum_{k=1}^I p_k \log_2 p_k \quad (8)$$

High value of uncertainty implies less similarity between users and vice versa. The similarity metric using the entropy (H) is formulated as

$$\text{SimE}(U_i, U_j) = 1 - \frac{H}{\log_2 I} \quad (9)$$

Considering (9) as the measure of similarity has a limitation. Consider a case when there is only one item that has been commonly rated between two users (while individually they have rated at least 20 items). In this scenario $\text{SimE}(U_i, U_j)$ will result in the similarity score of 1 even if there is high difference in the rating values of the two users for that item. $\text{SimE}(U_i, U_j) = 1$ implies that the users are highly similar which in reality does not hold true. To overcome this limitation, another factor of cosine similarity is added to the similarity metric as

$$\text{Sim}(U_i, U_j) = \beta \times \text{SimC}(U_i, U_j) + (1 - \beta) \times \text{SimE}(U_i, U_j) \quad (10)$$

where $\text{SimC}(U_i, U_j)$ is the cosine similarity score between users U_i and U_j .

In Eq. (10), β controls the dependency of similarity on each of the combined measure, determined using cross validation. If $\beta = 1$ similarity depends solely on cosine similarity measure and if $\beta = 0$ similarity depends solely on entropy based similarity measure. In order to have equal dependency on both similarity measures β is set to a value of 0.56 in our experiment. This step results in the set of similar users D to the target user U_i , filtered using a threshold on the similarity score.

Step 5: Item Prediction for target user

This is final step in the recommendation approach which computes the prediction coefficient of unrated item for the target user based on the similar users identified in the above step. The prediction coefficient $r_{U_i,k}$ for unrated item k is defined as

$$r_{U_i,k} = \frac{\sum_{s=1}^D r_{sk}}{|D|} \quad (11)$$

where D is the number of similar users to target user U_i and r_{sk} is the rating given by similar user $s \in D$ to the item k . Finally top N items are recommended to the user based on the value of prediction coefficient for the items.

3.2 Algorithm

The algorithm for the proposed approach is stated in Algorithm 1. Table 1 lists the symbols and their description used in Algorithm 1.

Table 1. Description of notations.

Symbols	Description
U_i	Target User
U_j	$j \in N - \{U_i\}$
G	Weighted Bipartite Graph
N	Number of Users
M	Number of items
E	Number of edges in a graph G
R	Sparse Rating Matrix of size $N * M$
C	Maximum number of co-rated items
r_{ui}	Rating of item i by user u
$p(U_i, U_j)$	Linear weighted sum for co-rated item filtering
$t(U_i, U_j)$	Percentage of co-rated items considered for similarity computation
$SimE(U_i, U_j)$	Entropy based similarity between users U_i, U_j
$SimC(U_i, U_j)$	Cosine Similarity measure between users U_i, U_j
$Sim(U_i, U_j)$	Similarity measure between users U_i, U_j
D	Number of similar users to U_i
I	Number of filtered co-rated items
Thresholduser_sim	Threshold for user similarity

Proposed Algorithm

Input: R, U_i

Output: Top n Items

Step1. Build a weighted bipartite graph $G = (N, M, E)$

Step2. Compute Degree of each user node of graph G .

Step3. Determine maximum number of co-rated items C between U_i and U_j
using (1). Let U_i be the user with minimum degree.

Step4. For each $k \in [1 \dots C]$

 Compute $|(r_{ik} - r_{jk})|$.

 count items with the corresponding absolute differences between ratings of the two users.

Step5. Compute $p(U_i, U_j)$ and $t(U_i, U_j)$ as in (2) and (3) respectively.

$I = \text{int}(p(U_i, U_j))$

Step6. Compute $SimE(U_i, U_j)$ and $SimC(U_i, U_j)$ using (9) and (11)
respectively.

Step7. Compute $Sim(U_i, U_j)$ as in (10)

Repeat steps 3-8 for all $j \in N - \{U_i\}$

Step8. $D =$ set of all users for which $Sim(U_i, U_j) > \text{Thresholduser_sim}$

Step9. Compute prediction coefficient for U_i using (12) and generate top n items.

4 Experiments and Results

This section presents the experiments and results used to evaluate the proposed approach. It details the dataset, the evaluation metric and the results of the experiments conducted. The Proposed Approach is compared with the traditional memory based Collaborative Filtering (CF) technique.

4.1 Dataset

Proposed Approach and CF approaches can work on any dataset. Performance of Proposed Approach and CF are evaluated using both public and private datasets.

MovieLens dataset is publicly available dataset which recommends movies to the user. The dataset consists of 100 K ratings (on a scale of 1–5) with 943 as the total number of users and 1682 movies. Each user rates at least 20 movies. Sparsity level of the dataset is 93.7% ($1 - (100000/(943 * 1682)) = 0.937$).

News dataset (Gautam et al. 2015) is privately available which recommends news items to the user. The dataset consists of 1 M ratings (on a scale of 1–5) that contain 100 users and 10 K news items. Each user rates 1 K news. Sparsity level of the dataset is 90% ($1 - (100000/(100 * 10000)) = 0.90$).

4.2 Evaluation Metric

Standard metrics in the area of information retrieval namely Precision (Herlocker et al. 2004), Recall (Herlocker et al. 2004) and F-Measure (Shani and Gunawardana 2011) are used to evaluate the performance of Proposed Approach and CF.

4.3 Results and Discussions

We compare Proposed Approach and CF in terms of performance. Results are tabulated in Tables 2 and 3 for both the techniques on the two different datasets.

Table 2. Results for MovieLens dataset.

Approach	Precision	Recall	F-measure
Proposed approach	0.109	0.953	0.195
CF	0.073	1.000	0.136

Table 3. Results for news dataset.

Approach	Precision	Recall	F-measure
Proposed approach	0.271	0.999	0.425
CF	0.174	1.000	0.296

Table 2 tabulates the results of the conducted experiment on MovieLens dataset. From Table 2 it is observed that the quality of recommendation improves for Proposed Approach as compared to CF. The ability to retrieve top ranked items that are mostly

relevant to the target user is increased by 49% in Proposed Approach as compared to CF. Recall for Proposed Approach decreased by 4.7% as compared to CF while F-Measure for Proposed Approach increased by 43% as compared to CF. Results show that the Proposed Approach is able to generate quality recommendation as compared to CF.

Table 3 tabulates the results for the conducted experiments on News dataset. For News dataset the Proposed Approach outperforms CF as is depicted in Table 3. Precision of Proposed Approach increased by 55% when compared with CF. Recall for Proposed Approach decreased by 0.1% as compared to CF and F-Measure for Proposed Approach increased by 43% outperforming CF. Results show that the Proposed Approach is able to generate quality recommendation compared to CF.

5 Conclusion

The paper presented a weighted bipartite graph based CF recommendation technique as the sparse nature of user-item rating matrix is easily represented in the form of a weighted bipartite graph to generate quality recommendations. After representing the rating matrix in the form of a weighted bipartite graph, the proposed approach then exploited the degree property of the graph to generate recommendations. To keep the spirit of collaborative filtering alive the proposed approach computes similarity between users using a hybrid similarity metrics which comprised of cosine similarity measure and a measure of information entropy. The proposed approach is compared with traditional CF approach for MovieLens and News datasets. Results showed that the proposed approach improved the overall performance of the system by 43% when compared with the CF approach on both the datasets.

Acknowledgement. The authors duly acknowledge Department of Computer Science, University of Delhi for extending their support and University Grants Commission (UGC) for funding this research work via Junior Research Fellowship (JRF) Ref No.: 3492/(NET-DEC 2012).

References

- Jannach, D.: *Recommender Systems: An Introduction*. Cambridge University Press, Cambridge (2010)
- Piao, C.-H., Zhao, J., Zheng, L.-J.: Research on entropy-based collaborative filtering algorithm and personalized recommendation in e-commerce. *Serv. Oriented Comput. Appl.* **3**(2), 147–157 (2009)
- Lee, K., Lee, K.: Escaping your comfort zone: a graph-based recommender system for finding novel recommendations among relevant items. *Exp. Syst. Appl.* **42**(10), 4851–4858 (2015)
- Huang, Z., Chung, W., Ong, T.-H., Chen, H.: A graph-based recommender system for digital library. In: *2nd ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 65–73 (2002)
- Chen, H., Gan, M., Song, M.: A graph model for recommender systems. In: *Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013)*, pp. 878–881 (2013)

- Huang, Z., Chen, H., Zeng, D.: Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. *ACM Trans. Inf. Syst.* **22**(1), 116–142 (2004)
- Sawant, S.: Collaborative filtering using weighted bipartite graph projection: a recommendation system for yelp. In: *Proceedings of the CS224W: Social and Information Network Analysis Conference* (2013)
- Chen, Y., Wu, C., Xie, M., Guo, X.: Solving the sparsity problem in recommender systems using association retrieval. *J. Comput.* **6**(9), 1896–1902 (2011)
- Lopes, R., Assunção, R., Santos, R.L.T.: Efficient Bayesian methods for graph-based recommendation. In: *10th ACM Conference on Recommender Systems*, pp. 333–340 (2016)
- Shams, B., Haratizadeh, S.: Graph-based collaborative ranking. *Exp. Syst. Appl.* **67**, 59–70 (2017)
- Hu, X., et al.: A hybrid recommendation model based on weighted bipartite graph and collaborative filtering. In: *IEEE/WIC/ACM International Conference on Web Intelligence Workshops (WIW)*, pp. 119–122 (2016)
- Wang, W., Zhang, G., Lu, J.: Collaborative filtering with entropy-driven user similarity in recommender systems. *Int. J. Intell. Syst.* **30**(8), 854–870 (2015)
- Chandrashekhar, H., Bhasker, B.: Personalized recommender system using entropy based collaborative filtering technique. *J. Electron. Commer. Res.* **12**(3), 214–237 (2011)
- Mehta, H., Bhatia, S.K., Bedi, P., Dixit, V.S.: Collaborative personalized web recommender system using entropy based similarity measure. *Int. J. Comput. Sci. Issues (IJCSI)* **8**(6), 231–240 (2011)
- Gautam, A., Radhika Dhingra, T., Bedi, P.: Use of NoSQL database for handling semi structured data: an empirical study of news RSS feeds. In: *Emerging Research in Computing, Information, Communication and Applications*, pp. 253–263 (2015)
- Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.* **22**(1), 5–53 (2004)
- Shani, G., Gunawardana, A.: Evaluating recommendation systems. In: *Recommender Systems Handbook*, pp. 257–297 (2011)

Automated Quiz Generator

Amit Bongir¹(✉), Vahida Attar¹, and Ramanand Janardhanan²

¹ College of Engineering Pune, Shivajinagar, Pune 411005, India
{bongiras13.it,vahida.comp}@coep.ac.in

² Choose To Think, Pune, India
ramanand@choosetothinq.com

Abstract. Automated Quiz Generator (AQG) is an extension of the factual question generation system implemented by Michael Heilman, which is generic and therefore applicable to any given domain of discourse in natural language. The extensions mainly include the ability to make MCQs out of generated questions and ranking questions by interestingness of the sentence in the input text from which the respective question was generated. Besides, it has functionality to extract interesting trivia from Wikipedia articles of important entities in the input text. Being domain independent, this system relies on DBpedia - a database of structured content extracted from Wikipedia, the largest general reference work on the Internet.

1 Introduction

1.1 MCQ Generation

Multiple-choice questions (MCQ) are arguably the most popular means of conducting objective tests. An MCQ comprises of:

- **Stem** – the question asked
- **Key** – the correct answer
- **Distractors** – incorrect alternatives to the key

Our system, the Automated Quiz Generator (AQG), passes a given, domain-independent piece of input text to Heilman's factual question generation system [Hei11] which provides a basic set of question-answer pairs which we regard as stem-key pairs of candidate MCQs. Our contribution is towards distractor generation, in which we make use of the Semantic Web technology - DBpedia [Leh+14]. Since DBpedia uses the Resource Description Framework (RDF) [LS99] data model to represent structured information extracted from Wikipedia, the methods we discuss throughout this report are applicable to any database adhering to the RDF specifications.

1.2 Ranking Generated Questions

Heilman's question generation (QG) system uses an *Overgenerate-and-Rank Framework* [HS09] for question generation. Due to overgeneration of questions,

ranking them according to their worth becomes necessary. Ranking in Heilman’s QG system is done mainly on linguistic features of generated question like grammaticality, length, pronoun replacement, etc. and each question is ranked according to the given score, which we’ll refer to as its *linguistic score*. In Sect. 3.4 we’ll discuss about ways to give an interestingness score to each sentence in the input text. The final score for ranking a question, which we’ll refer to as the question’s *value*, will be computed using both the question’s linguistic score and the interestingness score of the source sentence from which the concerned question was generated.

1.3 Trivia Extraction

Forbes is a renowned media company focussed on business, investing, technology, entrepreneurship, leadership, and lifestyle. Thus, the fact that the kind of readers that visit its website feel a nice welcome with a *Quote of the day* on every visit, works very well towards a critical economic aspect - user engagement. But quotes aren’t suitable to every kind of website. A more generic alternative is *Trivia* - interesting facts that are not well-known.

In this report, we discuss methods for extracting trivia about any DBpedia entity from the entity’s Wikipedia article. The novelty of our methods lies in the translation of the concepts introduced by Tsurel et al. [Tsu+16] befitting to an RDF database like DBpedia.

2 Related Work

2.1 MCQ Generation

There has been much work regarding MCQ generation from domain ontologies. Papasalouros et al. [PKK08] discuss ways of MCQ generation by selecting distractors using *class-based*, *property-based* and *terminology-based* strategies. OntoQue [AIY11] applies the strategies introduced by Papasalouros et al. [PKK08] and presents *fill-in the blank* type of questions as stem. Cubric and Tomic [CT11] optimize the strategies introduced by Papasalouros et al. [PKK08] and use them in their Protégé¹ [TC09] ontology editor. They further discuss annotation-based strategies using annotated information from domain ontology to generate stems and distractors. They present semantic similarity of distractors to the key as a combination of text and ontological similarity.

An important difference between these and other ontology-based techniques [Lop+15] and AQG lies in the method of constructing stem-key pairs. AQG accepts input text from the user, which acts as a syllabus pertaining to which stem-key pairs are generated using Heilman’s QG system [Hei11]. Whereas the

¹ <http://protege.stanford.edu/>.

other mentioned techniques generate stem-key pairs from the domain ontology and consider no syllabus to confine the generated MCQs to. These systems would require constructing a new ontological database or editing existing database to generate syllabus specific MCQs.

We present distractor generation of only entities that are proper nouns as proof-of-concept, since distractor generation techniques based on WordNet for entities that are common nouns have been previously explored [MH03].

2.2 Interestingness and Trivia Extraction

Tsurel et al. [Tsu+16] present the concepts of a category’s *similarity* to an article tagged by that category and its *cohesiveness*. These concepts were then used to find a category’s *trivia-worthiness* for a given article. Our work makes significant use of these concepts for trivia extraction as well as assigning interestingness score to input text sentences. Prakash et al. [Pra+15] propose their approach for automatically mining trivia from unstructured text, as they call it - “Wikipedia Trivia Miner (WTM)”. They make use of *unigram*, *linguistic* and *entity* features for their machine learning based *interestingness ranker*. The *unigram* and *entity* features are learnt with the training dataset as human voted trivia retrieved from the Internet Movie Database (IMDb) and as such the WTM is more accurate in extracting trivia related to movies.

3 System Design

The rest of the paper assumes that the reader is familiar with RDF [LS99] and the W3C standardized RDF query language - SPARQL. We elaborate on each of the stages depicted in Fig. 1 in the following sections. We also present important SPARQL queries used at various stages which could be run on DBpedia’s Virtuoso SPARQL Query Editor². The editor has predefined namespace prefixes³ and default data set name set to <http://dbpedia.org>. Therefore, we avoid specifying PREFIXes and FROM <<http://dbpedia.org>> in each query.

3.1 Important RDF Properties We Need

- `rdf:type` - denotes classes the entity is an instance of
Eg: `dbo:Scientist` is one `rdf:type` of `dbr:Alan_Turing`
- `dct:subject` - denotes topics related to the entity
Eg: `dbc:Fellows_of_the_Royal_Society` is a `dct:subject` of `dbr:Alan_Turing`

Henceforth we’ll refer to `rdf:type` and `dct:subject` as just **types** and **subjects** of an entity respectively.

² <https://dbpedia.org/sparql>.

³ <https://dbpedia.org/sparql?nsdecl>.

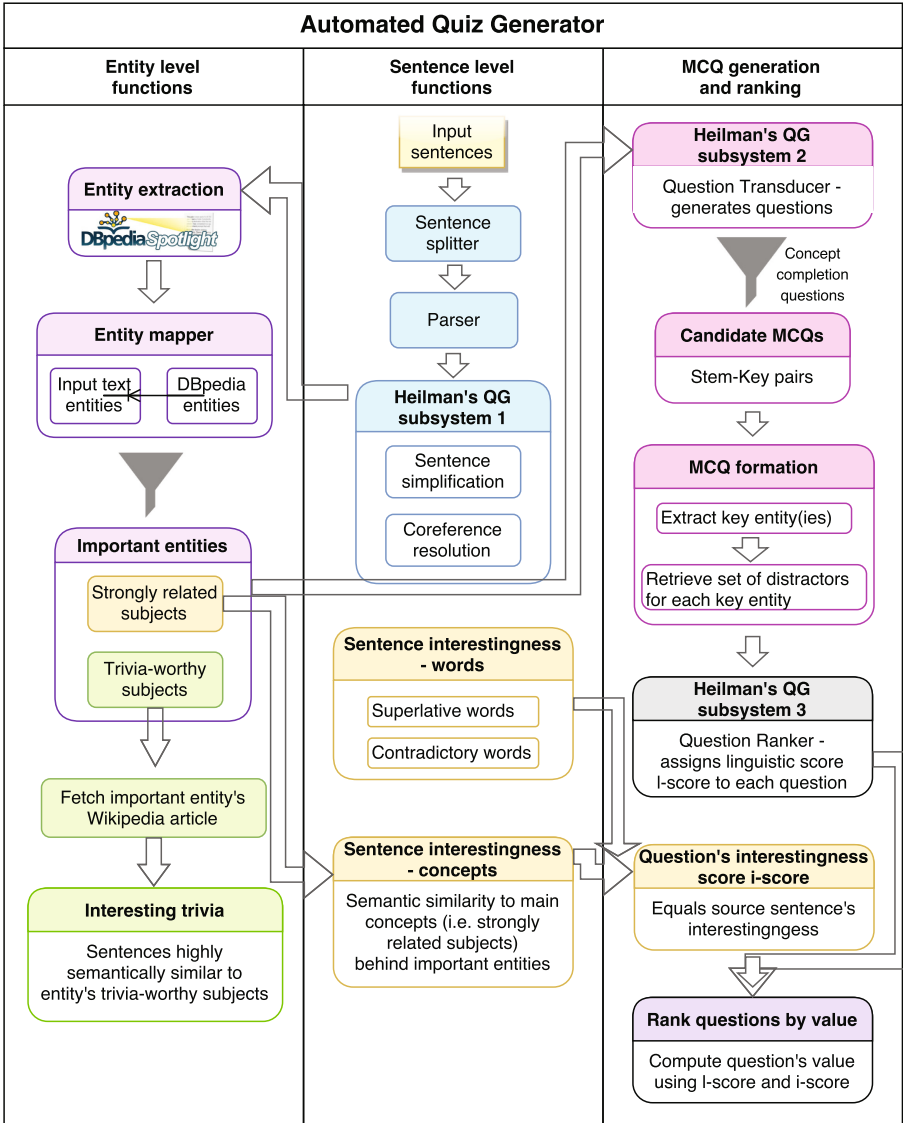


Fig. 1. Architecture of AQQ

`dc:subject` is one RDF property that has both forward and backward links [Ber06] in DBpedia. This allows retrieving all entities under a subject in one go, thus improving run-time performance. Every SPARQL query in the following sections has benefited from this aspect (Fig. 2).

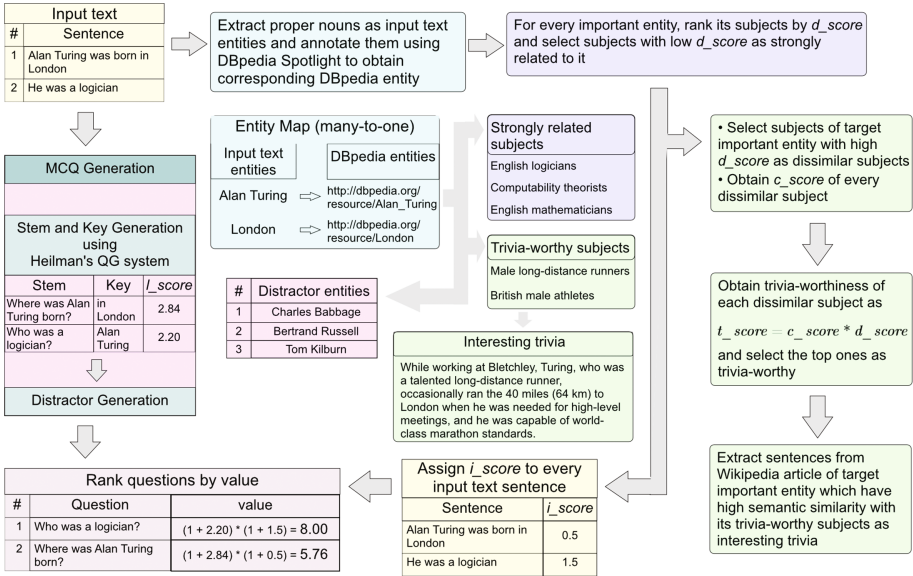


Fig. 2. Overview of AQG with input text related to Alan Turing as an example

3.2 Extracting DBpedia Entities from Input Text

The input text is first parsed by the Stanford Parser [K+03]. We then extract named entities in the input text and map them to their corresponding DBpedia entities. For each sentence in input text,

1. According to the TregexPattern⁴ [LA06] $NP < NNP \& !<< NP$, we extract noun phrases that dominate at least one proper noun and not any other noun phrase, as named entities
2. We apply fuzzy matching techniques to search for a named entity amongst the already mapped entities
3. If the search fails, we annotate the entire sentence using DBpedia Spotlight [Dai+13]. We pass the entire sentence to Spotlight so that it can apply Word Sense Disambiguation (WSD) and annotate even the following named entities in the sentence.

3.3 MCQ Generation

3.3.1 Stem and Key Generation

Heilman's question generation (QG) system generates 2 types of *sentence-level* factual questions [Hei11]:

⁴ TregexPattern javadoc page: <https://nlp.stanford.edu/nlp/javadoc/javanlp/edu/stanford/nlp/trees/tregex/TregexPattern.html>.

1. **Concept completion questions** - elicits a given partial concept or proposition
Eg: *Who served as the first Secretary of State under George Washington?*
2. **Verification questions** - yes-no answer
Eg: *Was Thomas Jefferson secretary of state?*

The ambitious reader interested in understanding the framework of the QG system is suggested to refer the work of Heilman and Smith [HS09].

We pass the given input text to Heilman’s QG system, from which we keep the obtained set of concept completion questions as candidate MCQs, whose question-answer pairs become the stem-key pairs. Heilman’s QG system assigns a score to each question based on its linguistic features like grammaticality, vagueness, pronoun resolution, etc., which we’ll refer to as its linguistic score l_score (Table 1).

Table 1. Question generation by Heilman’s QG system

#	Question	Answer	Source sentence	l_score
1	Where was Alan Turing born?	In London	Alan Turing was born in London	2.84
2	Who was a logician?	Alan Turing	He was a logician	2.20

3.3.2 Distractor Generation

The effectiveness of an MCQ is determined by the tendency of its distractors to be selected by the test-taker as an answer. This tendency signifies the distractor quality. An obvious measure of distractor quality is the semantic similarity between the key and concerned distractor. We obtain similarity between two DBpedia entities E_1 and E_2 as

$$sim(E_1, E_2) = |t(E_1) \cap t(E_2)|$$

where $t(E)$ denotes types of entity E . Using the entity map built in Sect. 3.2, we extract DBpedia entities in a candidate MCQ’s key, which we’ll refer to as *key entities*. For each key entity K , we construct our SPARQL query according to the following procedure:

1. Fetch all DBpedia entities under every subject of K as distractor entities
2. Rank the distractor entities in descending order of their similarity to K

Following is the SPARQL query for the key entity Alan Turing

```
SELECT ?distractor, (COUNT(DISTINCT ?type) AS ?similarity)
WHERE {
  dbr:Alan_Turing dct:subject ?subject .
  ?distractor dct:subject ?subject .
  FILTER (!SAMETERM(?distractor, dbr:Alan_Turing)) .
  dbr:Alan_Turing rdf:type ?type .
  ?distractor rdf:type ?type .
}
GROUP BY ?distractor
ORDER BY DESC(COUNT(DISTINCT ?type))
```

Therefore, for each key entity, we obtain the corresponding set of ranked distractor entities. We make use of a key entity’s set of distractor entities as a stack, circularly popping distractor entities every time as the key entity is found in keys of multiple candidate MCQs. The number of distractor entities we pop equals one less than the number of MCQ options desired; one left for the key itself.

We construct a complete distractor by replacing the key entity with the distractor entity. For a candidate MCQ whose key contains multiple key entities, we generate multiple MCQs with common stem-key pair, but with different sets of distractors according to each key entity.

3.4 Ranking Generated Questions

Summarization is a good strategy in Natural Language Processing (NLP) that provides a summary highlighting important points in an article. The intentions behind our strategy of scoring each sentence by interestingness are similar to those of summarization.

3.4.1 Sentence Interestingness and *i_score*

We first select important DBpedia entities in the input text. We say an entity to be important if its number of occurrences in the text is at least 50% of that of the most occurring DBpedia entity. One criteria of the interestingness of a sentence is then its semantic similarity to the main concepts behind each important entity. By main concepts, we refer to the strongly related subjects (Sect. 3.4.3) of a DBpedia entity. We use Wordnet-based techniques for computing semantic similarity of a sentence to a subject.

We also infer interestingness of a sentence by the presence of superlative words (like *first*, *best*, etc.) and contradictory words (like *but*, *although*, etc.) [Pra+15] (Tabel 2).

The interestingness score *i_score* of a sentence is then just the sum of the weights of its interestingness features shown in the above table. Note that a

Table 2. Sentence Interestingness Features

Feature	Weight
Semantic similarity to a strongly related subject of an important entity	0.25
Presence of one or more superlative words	0.50
Presence of one or more contradictory words	0.50

weight of 0.25 *each* is added for *every* strongly related subject of an important entity that is semantically similar to the sentence.

3.4.2 Computing Question's Value

The l_score of a generated question along with the interestingness score i_score (≥ 0) of the sentence in the input text from which the concerned question was generated is used to obtain the question's worth as,

$$value = \begin{cases} (1 + l_score) * (1 + i_score), & \text{if } l_score > 0 \\ l_score + i_score, & \text{otherwise} \end{cases}$$

Following are 2 of the ranked MCQs generated for the input text we have considered

Q1) Who was a logician?

1. Charles Babbage
2. Bertrand Russell
3. Alan Turing
4. Tom Kilburn

Answer: 3

Source sentence: He was a logician

Value: 8.00 ($l_score = 2.20$, $i_score = 1.5$)

Q2) Where was Alan Turing born?

1. in York
2. in London
3. in Boston, Lincolnshire
4. in Chichester

Answer: 2

Source sentence: Alan Turing was born in London

Value: 5.76 ($l_score = 2.84$, $i_score = 0.5$)

3.4.3 Strongly Related Subjects

The work of Tsurel et al. [Tsu+16] is based on extracting trivia-worthy categories of a Wikipedia article. In the context of DBpedia, an article becomes synonymous to a DBpedia entity and categories to the entity's subjects. Therefore, in accordance with the definition of similarity of an article to a category as discussed in [Tsu+16], we obtain the similarity of a DBpedia entity E to one of its subjects S_E as the average similarity between E and all DBpedia entities (except E of course) under S_E ,

$$\text{sim}(E, S_E) = \frac{\sum_{\substack{E_i \in S_E \\ E_i \neq E}} \text{sim}(E, E_i)}{|S_E| - 1}$$

The relevant SPARQL query for DBpedia entity **Alan Turing** is as follows,

```
SELECT ?subject, (STR(xsd:double(
  xsd:double(COUNT(DISTINCT ?entity))/
  xsd:double(COUNT(?type)))) AS ?d_score)
WHERE {
  dbr:Alan_Turing dct:subject ?subject .
  ?entity dct:subject ?subject .
  FILTER (!SAMETERM(?entity, dbr:Alan_Turing)) .
  dbr:Alan_Turing rdf:type ?type .
  ?entity rdf:type ?type .
}
GROUP BY ?subject
ORDER BY ASC(?d_score)
```

It ranks all subjects of **Alan Turing** according to their dissimilarity score d_score (≥ 0), which is nothing but inverse of the similarity score. Setting the maximum d_score as 110% of the d_score of the least dissimilar subject as cutoff, we select the strongly related subjects.

3.5 Trivia Extraction

We extract interesting trivia about only the important DBpedia entities in the input text (refer Sect. 3.4.1 for importance criteria). This is done by extracting sentences from the Wikipedia article of an important entity which are highly semantically similar to trivia-worthy subjects (Sect. 3.5.1) of the entity. We obtain the URL of the Wikipedia article of a DBpedia entity using its `foaf:primaryTopic` RDF property and therefore the article's text in response to a query sent to Wikipedia's `TextExtracts` API.

3.5.1 Trivia-Worthy Subjects

We make further use of the set of subjects ranked by dissimilarity of every important entity obtained in Sect. 3.4.3 to select trivia-worthy subjects. Setting the minimum d_score as 25% of the d_score of the most dissimilar subject as cutoff, we obtain dissimilar subjects $D_{E_{imp}}$ s of an important entity E_{imp} , which would be candidate trivia-worthy subjects for E_{imp} .

We then compute the cohesiveness score c_score of each $D_{E_{imp}}$. In accordance with [Tsu+16], we obtain cohesiveness of a dissimilar subject as the average similarity between pairs of entities in it. Instead of calculating similarity of entities pairwise, which would have a time complexity of $O(n^2)$, we implement a linear time algorithm as discussed below.

We first obtain the distinct types occurring across all entities under $D_{E_{imp}}$ excluding E_{imp} 's types,

$$t(D_{E_{imp}}) = (t(E_1) \cup t(E_2) \dots \cup t(E_m)) - t(E_{imp}) \mid E_1, \dots, E_m \in D_{E_{imp}}$$

We then keep a count of every type in $t(D_{E_{imp}})$ across all entities under $D_{E_{imp}}$. Setting the minimum count for cutoff as 1% of the number of distinct types $|t(D_{E_{imp}})|$, we exclude the types having low count. This signifies that the excluded type had a low contribution towards the average pairwise similarity of entities under $D_{E_{imp}}$. We divide the sum of the counts of the remaining types with $|t(D_{E_{imp}})|$ to get the cohesiveness score of $D_{E_{imp}}$.

$$c_score(D_{E_{imp}}) = \frac{\sum_{t_{sim} \in t(D_{E_{imp}})} \text{count}(t_{sim})}{\text{count}(t_{sim}) > 0.01 * |t(D_{E_{imp}})|} \mid |t(D_{E_{imp}})| + 1$$

We add 1 to $|t(D_{E_{imp}})|$ to handle the case $|t(D_{E_{imp}})| = 0$ which happens when $\forall E_i \in D_{E_{imp}}, t(E_i) \subseteq t(E_{imp})$.

The denominator consists of the number of *distinct types* instead of the number of *entities* under $D_{E_{imp}}$ for averaging since it provides a clearer idea of *incoherence* - more the variation in types across entities, greater is $|t(D_{E_{imp}})|$, and hence lesser the cohesiveness.

The relevant SPARQL query for obtaining cohesiveness of the dissimilar subject `Male long-distance runners of Alan Turing` is as follows,

```

SELECT (STR(xsd:double(SUM(?typeMatchCount))/
  xsd:double(?numEntityTypes) + 1) AS ?c_score)
WHERE {
  {
    SELECT (COUNT(DISTINCT ?type) AS ?numEntityTypes)
    WHERE {
      {
        ?entity dct:subject dbc:Male_long-distance_runners .
        ?entity rdf:type ?type .
      }
      MINUS { dbr:Alan_Turing rdf:type ?type . }
    }
  }
  {
    SELECT ?type (COUNT(?type) AS ?typeMatchCount)
    WHERE {
      {
        ?entity dct:subject dbc:Male_long-distance_runners .
        ?entity rdf:type ?type .
      }
      MINUS { dbr:Alan_Turing rdf:type ?type . }
    }
    GROUP BY ?type
  }
  FILTER (?typeMatchCount > 0.01 * xsd:double(?numEntityTypes)) .
}
GROUP BY ?numEntityTypes

```

Further, according to [Tsu+16], we obtain trivia-worthiness of a subject as,

$$t_score = c_score * d_score$$

The intuition behind this scoring of trivia-worthiness of a subject of an entity is that, while entities under the subject except the target entity are similar to each other, the target entity itself, on an average, is dissimilar to every other entity under the subject. Thus, it follows that the concerned subject is unexpected/-surprising for the target entity, and hence trivia-worthy.

We set the minimum t_score for trivia-worthy subjects as 75% of the t_score of the most trivia-worthy subject.

4 Evaluation

4.1 MCQ Generation

We evaluated MCQ generation with input data set as 15 documents related to 15 diverse fields of knowledge, each document consisting of short descriptions about

3 renowned entities in the respective field. Each MCQ is generated with 4 options. In each document, we select only those MCQs for evaluation having a value of at least 40% of that of the most valued question. Further, from MCQs having common stem-key pair, we select the one having the best set of distractors. Thus, a total of 140 MCQs were selected for evaluation.

The worth of distractors in an MCQ varies largely based on the knowledge of the test-taker about the topic on which the MCQ is based. An expert would directly know the answer and hence distractors would be useless. Therefore we take a comparative approach in evaluating the quality of distractors. We categorized each MCQ based on its distractors into one of the following 4 types (Table 3):

1. **Good** – highly semantically similar to the key
2. **Wrong due to external factors** – external factors mainly include:
 - wrong coreference resolution by Heilman’s QG system
 - wrong entity recognition by DBpedia Spotlight
 These errors led to obtaining distractor entities for the wrong entity, and consequently wrong distractors.
3. **Obviously wrong** – these are ones which require no expertise by the test-taker but can be simply eliminated based on metadata provided by the stem. For example - an MCQ which requires a female person as its answer. Presence of words like “she”, “her”, etc. in its stem provide the hint that the answer is a woman. A male person as a distractor in this case would obviously be wrong.
4. **Another answer** – a distractor that itself is another answer to the question

Table 3. MCQs categorised based on distractors

Category	Percent
Good	57.14
Wrong due to external factors	24.29
Obviously wrong	15.00
Another answer	3.57

The set of distractors of an MCQ is labeled *good* if all distractors in it are so, otherwise it gets labeled by one of the other 3 categories if even one of the distractors belongs to it. The last 2 categories - *obviously wrong* and *another answer* - are due to the flaw that we haven’t considered the context of the question while preparing distractors. Reducing them would require methods mainly pertaining to the field of question answering, which were beyond the limitations of this system.

It should be noted that MCQs categorised as *good* don’t ensure that they are fit to be used directly in a test. It was observed that **48.75%** of the *good* MCQs required *post-editing*. By post-editing, we mean

- making the MCQ stem/options grammatically correct
- reducing vagueness of the stem
- reducing length of the options by removing unnecessary information repeating in every option. Since distractor generation involved simply replacing key entity with distractor entity, the information surrounding the key entity gets repeated in every option, which could be removed and instead be used to make the stem more informative, hence reducing its vagueness.

Such post-editing doesn't involve modifying distractor entities themselves.

Thus, it is evident from the results that our MCQ generation methods provide assistance rather than replacement to a test-setter.

4.2 Trivia Extraction

We evaluate trivia extracted by AQG for the same entities belonging to 15 diverse fields of knowledge we used for evaluating MCQs. 1–2 facts per entity were generated, making a total of 81 facts for evaluation. The trivia were categorised into one of the following categories and subcategories (Tabel 4):

1. Interesting

- (a) *Surprising* – a fact that is unexpected to be known of an entity
- (b) *Just didn't know before* – a fact about an entity that may not be so well-known but not surprising for an entity. For example, the fact that “*Mike Tyson was the first heavyweight boxer to simultaneously hold the WBA, WBC, and IBF titles*” may not be well-known but not surprising for a legendary boxer like Mike Tyson. But it is surely surprising that “*To help pay off his debts, Tyson returned to the ring in 2006*”.

2. Not interesting

- (a) *Lame*
- (b) *Knew before*

3. Couldn't understand

- (a) *Vague* – caused by AQG's attempt to keep the trivia as short as possible
- (b) *Was about wrong entity* – caused by the same problems of wrong coreference resolution by Heilman's QG system and wrong entity resolution by DBpedia Spotlight that troubled MCQ generation

Table 4. Results for trivia extraction

Category	Percent	Constituents	
		Type	Percent
Interesting	49.38	Surprising	45.00
		Just didn't know before	55.00
Not interesting	38.27	Lame	45.16
		Knew before	54.84
Couldn't understand	12.35	Vague	60.00
		Was about wrong entity	40.00

In terms of accuracy in obtaining trivia-worthy subjects, AQG has similar performance to Fun Facts [Tsu+16]. The extracted trivia couldn't be compared since Fun Facts doesn't extract the trivia sentence, as we do from the entity's Wikipedia article. The crucial improvements in performance were observed in terms of execution time. While Fun Facts took about 1.45h to extract trivia-worthy categories for 2 entities - **Barack Obama** and **Bill Clinton** - AQG obtained the top trivia-worthy category **Grammy Award winners** as obtained by Fun Facts for both entities and extracted the concerned trivia sentences from their respective Wikipedia article in about 3.5 min on the same machine.

5 Future Work

The common reasons that affected both MCQ generation and trivia extraction were external factors like wrong coreference resolution by Heilman's QG system and wrong entity recognition by DBpedia Spotlight. Heilman's QG system uses ARKref [OH13] for coreference resolution, and thus AQG will benefit from its improvements. DBpedia Spotlight provides a ranked list of n -best candidate DBpedia entities as annotations for a named entity in the input text. Currently, we simply select the top most ranked annotation for every input text entity. In doing so, we are being completely dependent on Spotlight's Word Sense Disambiguation (WSD) techniques it applies to the sentence we pass each time we want to get an entity in that sentence annotated. Rather, we could apply WSD techniques on the entire input text natively first, and then select the best candidate amongst the ranked annotations provided by Spotlight. There is also a correlation between coreference resolution and entity recognition by Spotlight that needs to be worked upon:

- Information about recognized entities can be used for coreference resolution
- More the number of coreferences resolved in a sentence, more is the information about the resolved entities available to Spotlight for WSD

Another important improvement would be considering context of an MCQ's stem for distractor selection. In this way, we'll not only be able to reduce the number of distractors that are *obviously wrong* or *another answer*, but also select distractor entities that are not just semantically similar to the key entity but also semantically related to the stem.

Shortening an answer to a question generated by Heilman's QG system to retain prominently the key entity in it and using the removed information to make the question more informative will reduce both vagueness of the question as well post-editing.

Tsurel et al. [Tsu+16] have suggested *personalization* of trivia. This is very much needed given the fact that interestingness of a trivia is highly subjective and depends on various parameters specific to a reader.

References

- [AlY11] Al-Yahya, M.: Ontoque: a question generation engine for educational assessment based on domain ontologies. In: 2011 11th IEEE International Conference on Advanced Learning Technologies (ICALT), pp. 393–395. IEEE (2011)
- [Ber06] Berners-Lee, T.: Backward and Forward links in RDF just as important (2006). <http://dig.csail.mit.edu/breadcrumbs/node/72>. Accessed 24 Apr 2017
- [CT11] Cubric, M., Tasic, M.: Towards automatic generation of e-assessment using semantic web technologies. *Int. J. e-Assess.* (2011)
- [Dai+13] Daiber, J., Jakob, M., Hokamp, C., Mendes, P.N.: Improving efficiency and accuracy in multilingual entity extraction. In: Proceedings of the 9th International Conference on Semantic Systems, pp. 121–124. ACM (2013)
- [Hei11] Heilman, M.: Automatic factual question generation from text, Ph.D. thesis, Carnegie Mellon University (2011)
- [HS09] Heilman, M., Smith, N.A.: Question generation via overgenerating transformations and ranking. Technical report, DTIC Document (2009)
- [K+03] Klein, D., Manning, C.D., et al.: Fast exact inference with a factored model for natural language parsing. In: *Advances in Neural Information Processing Systems*, pp. 3–10 (2003)
- [LA06] Levy, R., Andrew, G.: Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, pp. 2231–2234. Citeseer (2006)
- [Leh+14] Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., et al.: DBpedia—a large-scale, multilingual knowledge base extracted from Wikipedia. *Semant. Web J.* (2014)
- [Lop+15] Lopetegui, M.A., Lara, B.A., Yen, P.-Y., Çatalyürek, Ü.V., Payne, P.R.: A novel multiple choice question generation strategy: alternative uses for controlled vocabulary thesauri in biomedical-sciences education. In: *AMIA Annual Symposium Proceedings*, vol. 2015, p. 861. American Medical Informatics Association (2015)
- [LS99] Lassila, O., Swick, R.R.: Resource description framework (RDF) model and syntax specification (1999)
- [MH03] Mitkov, R., Ha, L.A.: Computer-aided generation of multiple-choice tests. In: *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing*, vol. 2, pp. 17–22. Association for Computational Linguistics (2003)
- [OH13] O’Connor, B., Heilman, M.: ARKref: a rule-based coreference resolution system (2013). arXiv preprint [arXiv:1310.1975](https://arxiv.org/abs/1310.1975)
- [PKK08] Papasalouros, A., Kanaris, K., Kotis, K.: Automatic generation of multiple choice questions from domain ontologies. In: *e-Learning*, pp. 427–434. Citeseer (2008)
- [Pra+15] Prakash, A., Chinnakotla, M.K., Patel, D., Garg, P.: Did you know?-mining interesting trivia for entities from wikipedia. In: *IJCAI 2015*, pp. 3164–3170 (2015)
- [TC09] Tasic, M., Cubric, M.: SeMCQ-Protégé Plugin for automatic ontology-driven multiple choice question tests generation. In: *Proceedings of the 11th International Protege Conference*. Stanford Center for Biomedical Informatics Research (2009)
- [Tsu+16] Tsurel, D., Pelleg, D., Guy, I., Shahaf, D.: Fun facts: automatic trivia fact extraction from wikipedia (2016). arXiv preprint [arXiv:1612.03896](https://arxiv.org/abs/1612.03896)

Temporal Modelling of Bug Numbers of Open Source Software Applications Using LSTM

Jayadeep Pati^(✉), Krishnkant Swarnkar, Gourav Dhakad, and K.K. Shukla

Indian Institute of Technology (BHU), Varanasi, Varanasi 221005, India
{jayadeep.rs.cse12,krishnkant.swarnkar.cse15,gourav.dhakad.cse15,
kkshukla.cse}@iitbhu.ac.in

Abstract. Predicting the number of bugs in any software application is an important but challenging task. The software manager by modelling the bug numbers, can take timely decisions in reducing the amount of effort investment and also the allocation of resources. The software developers can also take effective steps for reducing the number of bugs in the future version of the software application. The end users also can make a timely decision on adoption of a particular software application by knowing the growth pattern of bugs in advance. The challenges behind modeling the bug growth patterns are random causes behind a bug. A bug in any software may be caused during testing, development or application. Causal modelling of bug numbers is a complex and tedious task as they consider many internal characteristics to be modelled. In this paper, we have used we have used Long Short Term Memory (LSTM) [14] Network for temporal modelling the bug numbers of three different software applications. We have used both univariate and multivariate modelling approach to predict bug number in advance. The goal is to have an appropriate model for software bug growth pattern.

1 Introduction

The likelihood of bugs in any software application depends upon many invariant parameters like code complexity, problem domain, amount code change and also on internal software process adopted [1]. A bug or defect may be caused during a different stage of development of software like coding, design or testing phase. The causal modelling of the bug growth pattern in any software is a complex and tedious task as it inquiries many internal details of software which sometimes is also impossible. The event of reporting a bug, fixing a bug and a new developer assigned to a project are all uncertain in advance [2]. However in aggregate, all these random like interactions shows some rules and patterns, which is not purely random. Effective time series modelling approach is required to model these bug growth pattern.

Previous studies on modelling of bug growth patterns are based on linear ARIMA model [3], where the author used to predict a stationary time series data. In another paper [4], polynomial regression technique is used to model

the temporal bug patterns in the Eclipse. The Nonlinear Autoregressive neural network is also used in one paper [5] for prediction of bug numbers in the Debian operating system. Some papers [6] also used a combination of the linear and nonlinear model to predict the bug growth patterns.

In this paper, we have used Long Short Term Memory (LSTM) Network [12] for modelling the increases and decreases in bug numbers. In the first model, we have used univariate modelling of bug number series. In the second model, we have used lag values of corresponding bug number series as well as a lag value other associated series into consideration. In the third model, we have used an additional factor covariance between data values for multivariate modelling the bug number series. The bug information for different software applications is kept in bug repositories [5]. We have extracted the bug number data from Debian Bug Repository [7], Eclipse Bug Repository [8] and Mozilla bug repository [9]. The bug number data are collected from Jan 2005 to Nov 2016. We split the yearly bug number data into monthly bug number data i.e. 155 months in sum. So we get three bug number series of length 155 each.

2 Long Short Term Memory (LSTM) Network Modelling

In this paper, we have used Long Short Term Memory (LSTM) Network [10] for modelling the increases and decreases in bug numbers. LSTM is a special variant of Recursive Neural Networks (RNN) [11]. Recurrent Neural Network [15] gives a bigger edge over other types of Neural Networks like Feedforward Neural Network. Each unit of RNN stores previous state and calculates the new state based on its previous states and the current inputs provided to it.

Let $h(t)$ denote the internal state of an RNN unit, $h^l(t)$ denote the internal state of the previous layer RNN unit (input), and W be the weight matrix of dimension, $n \times 2n$, then: $h(t) = \tanh\left(W \times \begin{bmatrix} h^l(t) \\ h(t-1) \end{bmatrix}\right)$ [12].

The Diagrammatic representation of RNN is given in Fig. 1. In practice RNN's work very efficiently when the gap in the dependencies is short. But as the gap grows RNN's seem to become less effective. This problem can be overcome using LSTM Networks.

The LSTM Model [13] uses hyperbolic tangent (\tanh) as the activation function. The sigmoid function is used as the gating function for the three (in, out and forget) gates:

$$\tanh(x) = \frac{(e^x - e^{-x})}{(e^x + e^{-x})}$$

C_t, h_t, x_t denotes the cell state, hidden state, input vectors respectively, $W_{xi}, W_{hi}, W_{xo}, W_{ho}, W_{xf}, W_{hf}, W_{xg}, W_{hg}$ are weight metrics and B_i, B_o, B_f, B_g are bias vectors [12].

$$\begin{aligned}
i_t &= \text{sigmoid}(W_{xi} \cdot x_t + W_{hi} \cdot h_{t-1} + B_i) \\
f_t &= \text{sigmoid}(W_{xf} \cdot x_t + W_{hf} \cdot h_{t-1} + B_f) \\
g_t &= \text{tanh}(W_{xg} \cdot x_t + W_{hg} \cdot h_{t-1} + B_g) \\
C_t &= f_t * C_{t-1} + i_t * g_t \\
h_t &= o_t * \text{tanh}(C_t) \\
o_t &= \text{sigmoid}(W_{xo} \cdot x_t + W_{ho} \cdot h_{t-1} + B_o)
\end{aligned}$$

i_t : Input gate controls how much the input vector and the hidden state vector changes the cell state.

o_t : Output gate controls how much the cell state affects the output of the cell.

f_t : Forget gate allows the cell to remember or forget the previous states.

g_t : Candidate value vector helps to calculate the cell state.

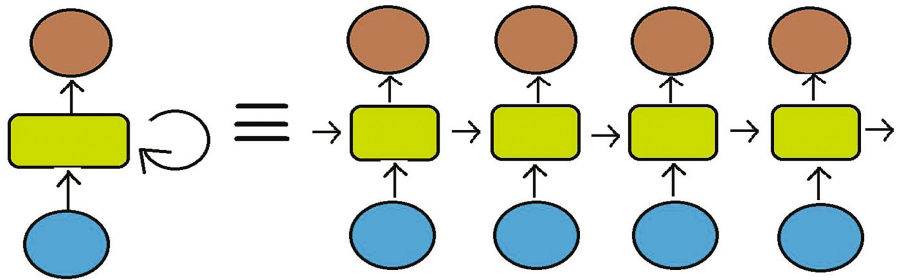
The mathematical insights of LSTM are given in Fig. 1. Figure 1 also presents the values of i , o and f gates. Figure 1 shows the structure of neural unit of LSTM Network.

2.1 Dataset Description

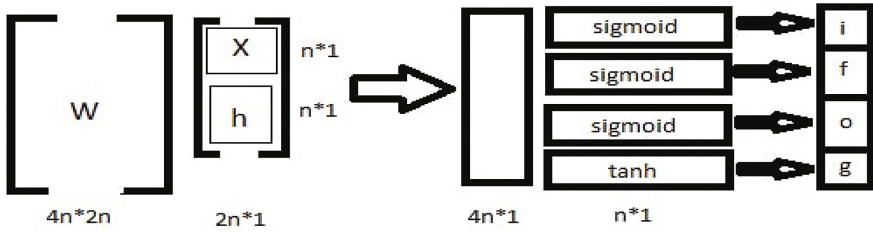
In this paper, we have analysed the bug growth pattern of three different software applications. We obtain the bug number data from public bug repository of Debian, Eclipse, and Mozilla respectively. The Debian Bug number data is available in the Debian Bug Repository in the Ultimate Debian Database (UDD) [7]. The Eclipse Bug number data is available in the Eclipse Bug Repository [8]. The Mozilla bug data is available in the Bugzilla [9].

Debian is an operating system for both stand-alone PC and servers. Debian uses the Linux kernel Gnu/Linux based OS tools and also is an important Linux distribution. We collected the bug information of Debian from January 2005 to November 2016 over 12 years. The total count of bug number is 62,563. We have divided the bug number data into monthly bug number data, that is, 155 months in sum.

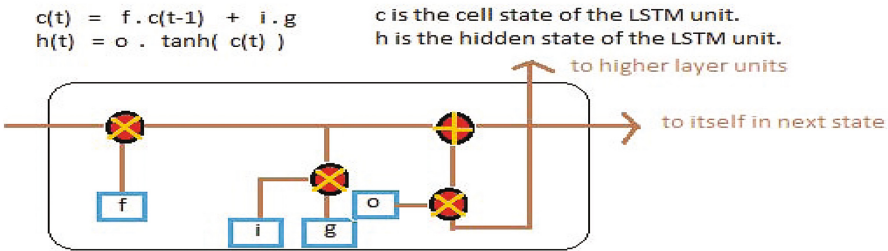
Mozilla is established in 1998 by members of NetScpae as an open source community. It has launched many products like Firefox browser, Firefox Mobile web browser, Firefox mobile operating system and other projects. The bug information is stored in Bugzilla bug tracking system. We also collected the bug number data for Mozilla from January 2005 to January 2015 with a total count of bugs as 2252. Similarly, the bug numbers data is divided into monthly bug number data, that is, 155 months in sum.



RNN Representation Diagram



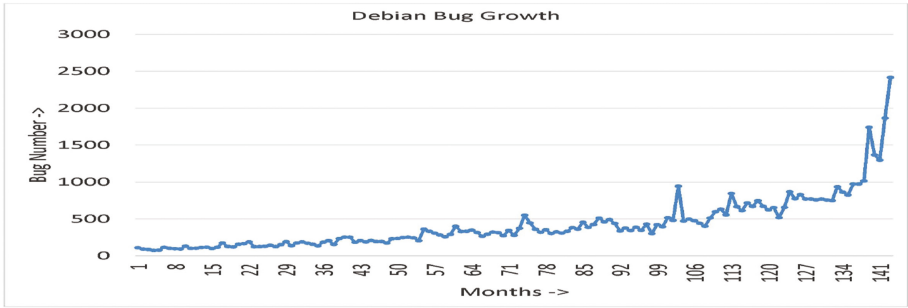
LSTM Mathematical Representation Diagram



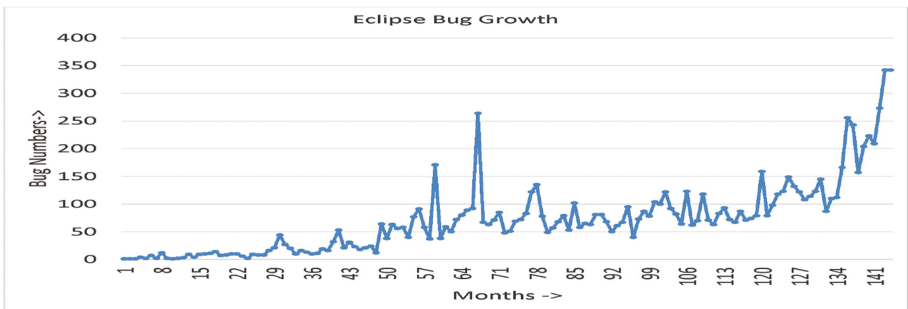
LSTM Structural Representation Diagram

Fig. 1. Diagrammatic representation: RNN & LSTM

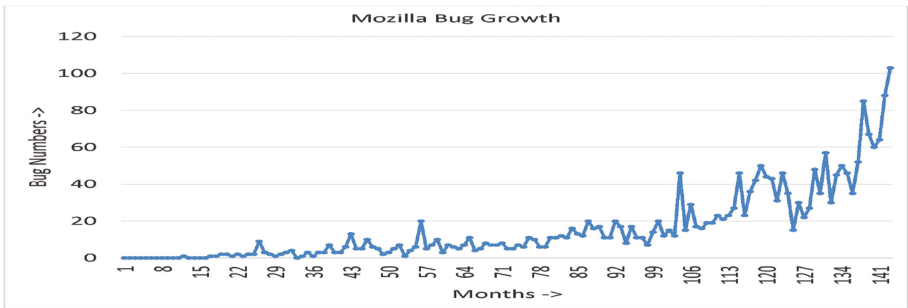
Eclipse is an important integrated development environment (IDE) frequently used in computer programming. It can also be used as an extensible plug-in with other environments. It is developed using Java and also primarily used in developing Java application. The Eclipse Project has developed Eclipse Bugzilla for bug tracking. The bug numbers data is divided into monthly bug number data, that is, 155 months in sum. The monthly time series of bug numbers for Debian, Mozilla, and Eclipse is given in Fig. 2 respectively.



Debian Bug Growth Pattern



Eclipse Bug Growth Pattern



Mozilla Bug Growth Pattern

Fig. 2. Diagrammatic representation of bug growth patterns

3 The Design of Experiments

We have designed three types of LSTM models for predicting the increases and decreases in bug numbers. In the first model, we have used univariate modelling

of bug number series. In the second model, we have used lag values of corresponding bug number series as well as a lag value other associated series into consideration. In the third model, we have used an additional factor covariance between data values for multivariate modelling the bug number series.

3.0.1 First Model (LSTM Univariate Modelling)

In univariate modelling, we have used the lag values of corresponding bug number series for modelling. We have applied univariate modelling to Debian, Eclipse, and Mozilla bug number Data. Here, there are three separate models for three bug number series. For example, for the Debian bug number series, we only took data from Debian Bugs Data (2 lags). The correlation between the data values is not taken into consideration. The LSTM network is implemented in the Python environment. We have used Python Keras Library [16]. Our model contains 1 hidden layer with 4 neural units. The lag values are calculated for the corresponding bug number series by Partial Auto-Correlation Function (PACF) plots. The Model Diagram is given in Fig. 3.

3.1 Second Model (LSTM Multi-variate Modelling)

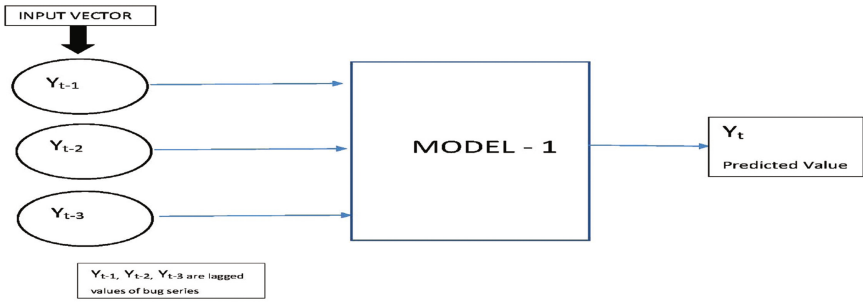
Here we have used lag values of corresponding bug number series as well as a lag value other associated series into consideration. This is called multivariate modelling as the effect of other bug number series are also taken into account. For example, for modelling Debian bugs, we took the data from Debian Bugs Data (2 lags) and also Mozilla Bugs (1 lag) into consideration. The lag values are calculated for the corresponding bug number series by Partial Auto-Correlation Function (PACF) plots. The Model Diagram is given in Fig. 3.

3.2 Third Model (LSTM Multi-variate Modelling with Covariance)

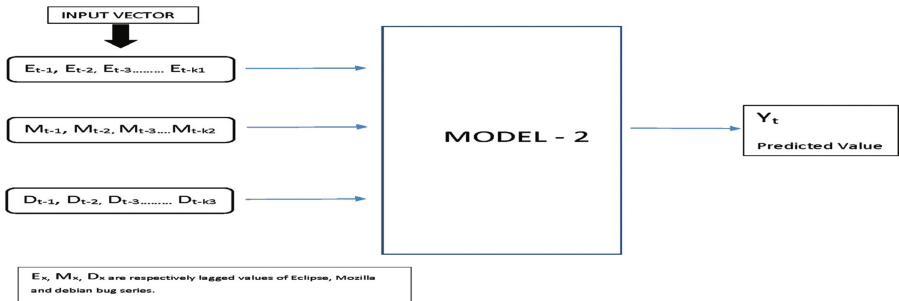
Here we have used lag values of corresponding bug number series, other associated series along with covariance between the data values of series into consideration. This is called multivariate modelling as the effect of other bug number series are also taken into account. For example, for modelling Debian bugs, we took the data from Debian Bugs Data (2 lags), Mozilla Bugs (1 lag), Covariance (2 length series of Mozilla and Debian Data), Covariance (2 length series of Debian and Debian Data). The lag values are calculated for the corresponding bug number series by Partial Auto-Correlation Function (PACF) plots. The Model Diagram is given in Fig. 3.

Covariance is the measure of how much two random variables vary with respect to each other. A High positive value of covariance indicates that if one of the random variable increases then other also increases, a High negative value of covariance indicates that if one of the random variables decreases then another one also decreases.

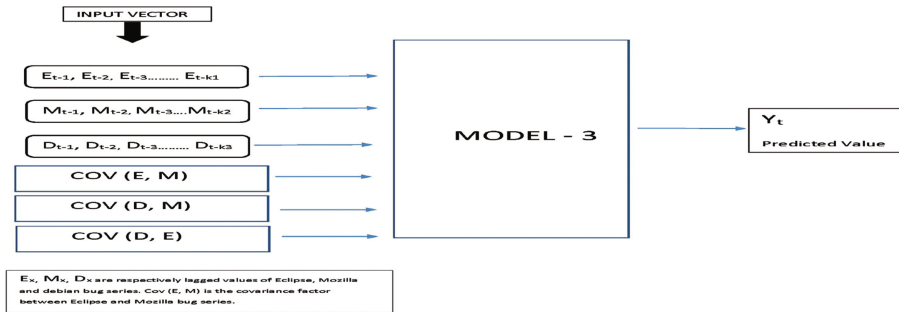
The LSTM network is implemented in Python environment. We used the Keras library in Python to model the bug number series. The detailed procedure of LSTM univariate and multivariate modelling is given below.



Diagrammatic Representation: Model 1



Diagrammatic Representation: Model 2



Diagrammatic Representation: Model 3

Fig. 3. Diagrammatic representation of LSTM models

Steps for Implementation of LSTM Univariate and Multivariate Modelling:

1. After getting the Bug Number Series of three different open source software applications, the next phase to model the bug number series with LSTM univariate and multivariate modelling.
2. First, we divided it into two parts 80% for training and rest 20% for testing.
3. The next step is to create an LSTM neural network and also to fix the number of initial layers.

4. The number of input to LSTM network is decided by the lag values of bug number series.
5. The covariance of the time series being predicted with the other time series is also taken as input (for Multivariate modelling).
6. The next step is to train the LSTM network on the training data.
7. Then Evaluation of the performance of the model on the test data is obtained on the basis of RMSE.

The Flow diagram of an implementation of LSTM is given in Fig. 4.

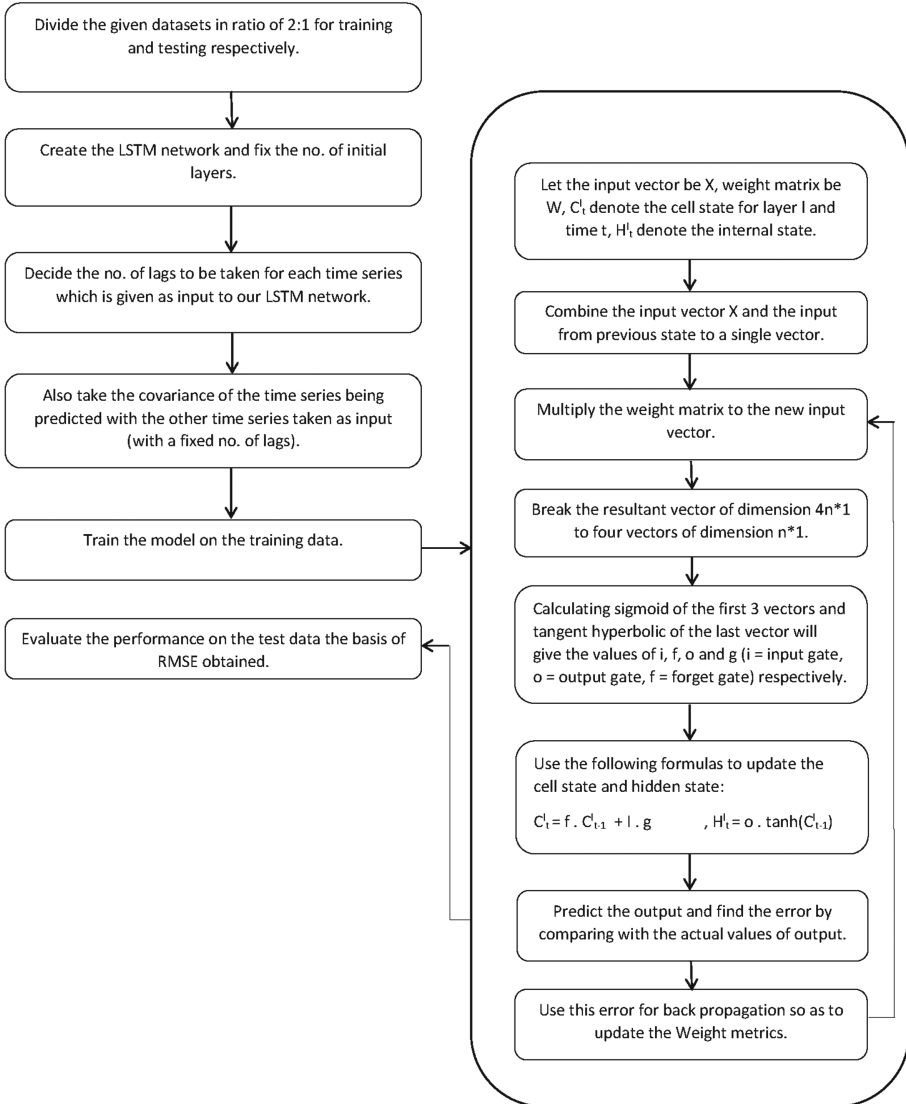


Fig. 4. LSTM model implementation diagram

4 Evaluation and Interpretation

The models are evaluated on the basis of Root Mean Square Error (RMSE) value as provided by the model on the train and test data. The model which gives better result on test data is selected. The RMSE is calculated by this formulae.

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^N (A(t) - F(t))^2}$$

Here: A(t): Actual Value, F(t): Predicted Value, N: Number of Terms.

First, the models are trained with LSTM network with the corresponding trained data (80%) for three bug number series. Then the trained models are tested with corresponding test data values (20%). The model which given minimum RMSE for test Data is considered as the most suitable model. The RMSE values for the test data as given by the three models are shown in Table 1. From the table, we observe that the RMSE value for Multivariate LSTM model is less than univariate LSTM model. This is because the Multivariate model considers also consider the interrelation between the time series into consideration. In the third model, we have also added the covariance factor into account. We observe that after adding the covariance factor the RMSE value further decreases. Covariance can be considered as an additional factor which improves the accuracy of the model. So we observed that the Multivariate LSTM model with covariance factor is the most suitable model for bug prediction. We have also presented the bar chart representation of RMSE value as given by three models in Fig. 5. We have also plotted the predicted series and original series for three different bug number series. Figure 6 represents the plots for Eclipse. Figure 7 represents the plots for Mozilla. Figure 8 represents the plots for Debian.

Table 1. RMSE value: test data

	Univariate (LSTM)	Multivariate (LSTM)	Multivariate covariance (LSTM)
Debian	252.5	230.77	210.86
Mozilla	16.3	15.71	14.92
Eclipse	79.81	77.03	67.44

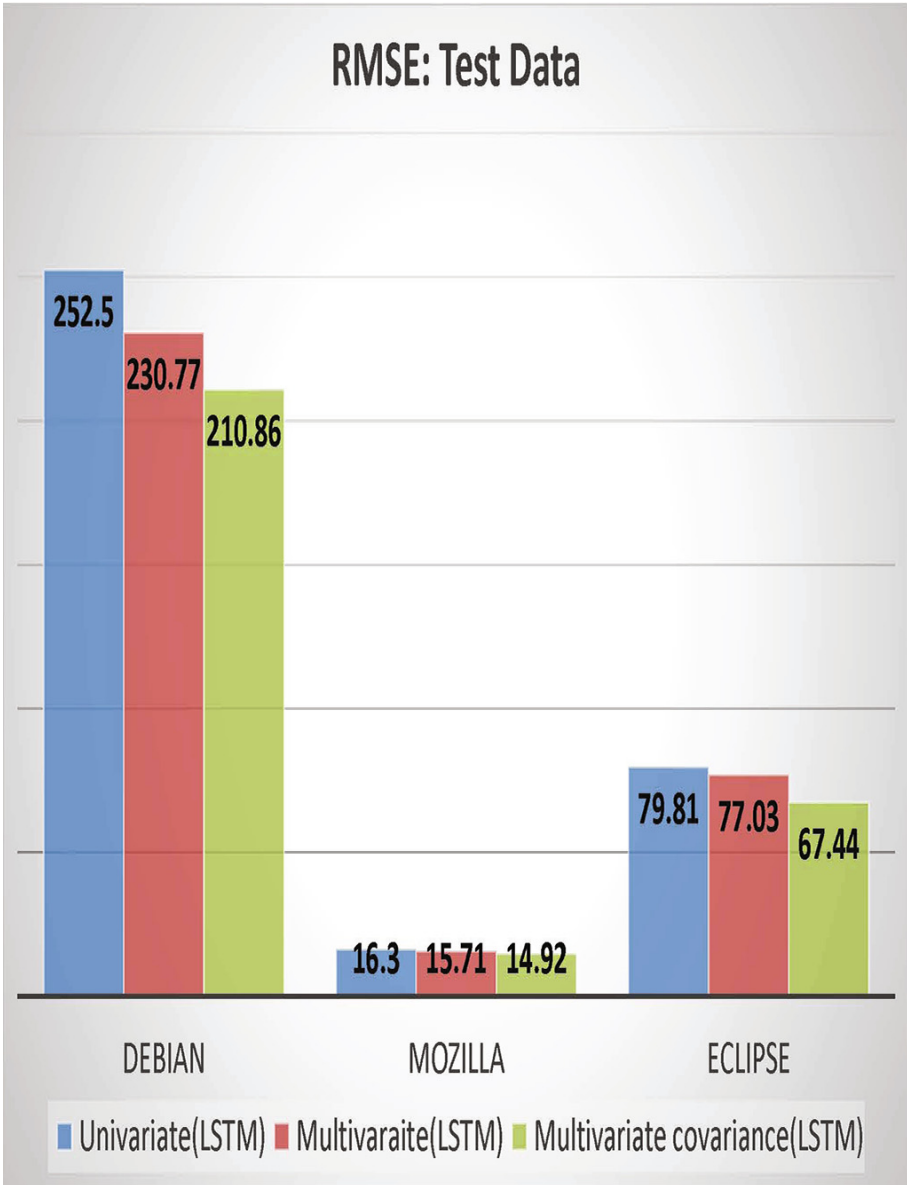
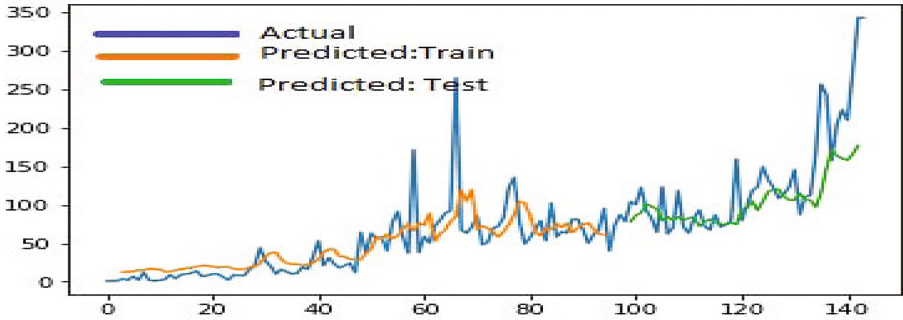
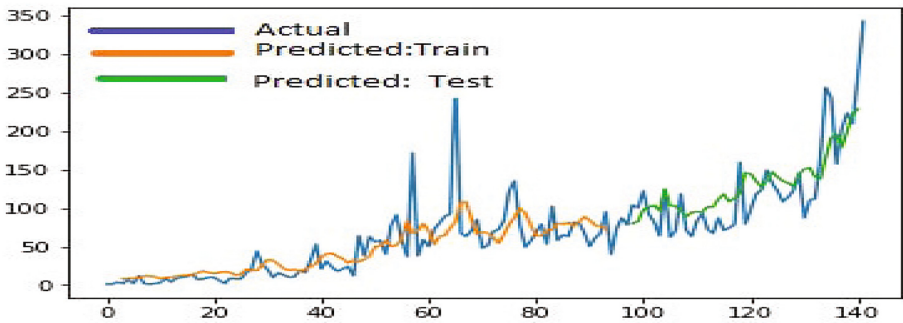


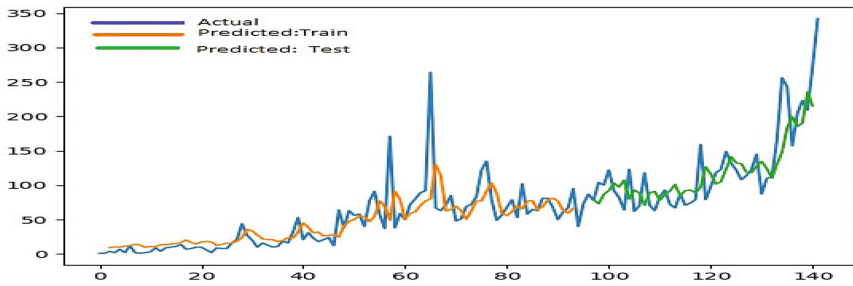
Fig. 5. RMSE comparison between models: Test data



Plots of Predicted Vs. Actual Bug Number Series (Univariate LSTM)

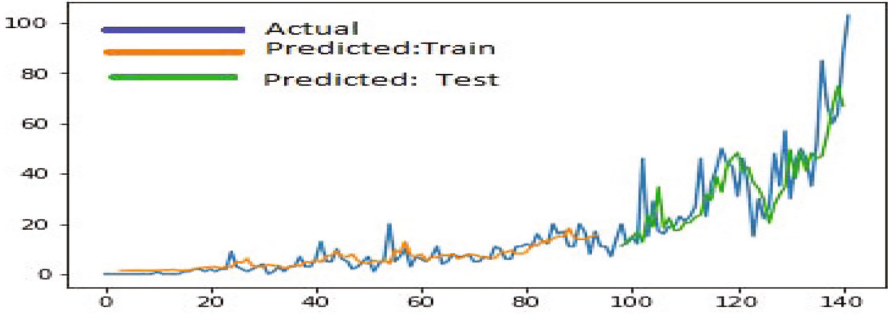


Plots of Predicted Vs. Actual Bug Number Series (Multivariate LSTM)

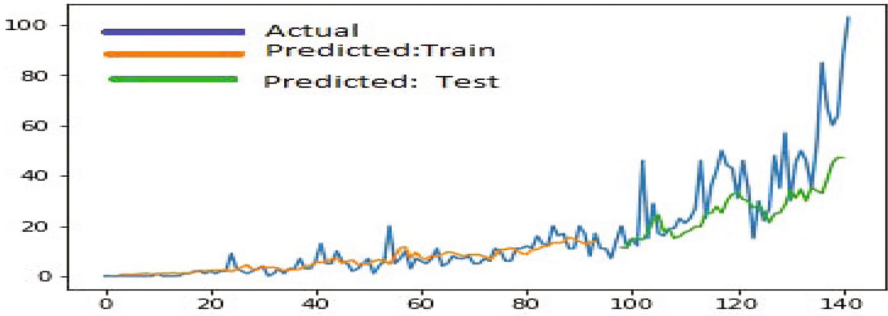


Plots of Predicted Vs. Actual Bug Number Series (Multivariate LSTM+ Covariance)

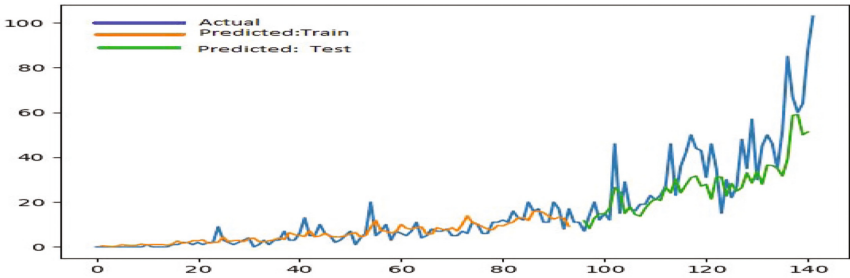
Fig. 6. Plots of predicted vs. actual bug number series: Eclipse



Plots of Predicted Vs. Actual Bug Number Series (Univariate LSTM)

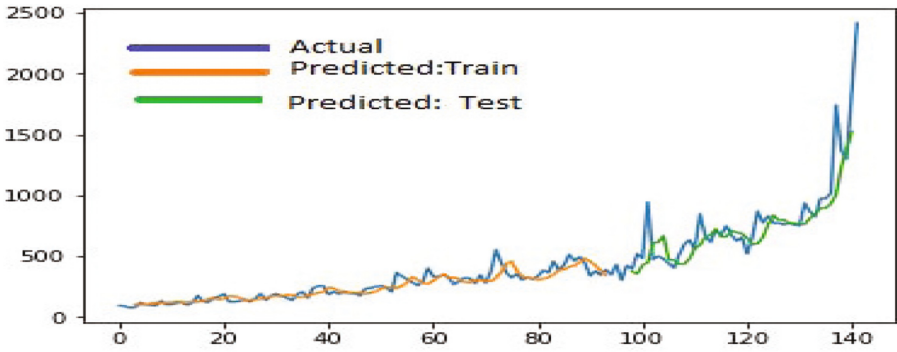


Plots of Predicted Vs. Actual Bug Number Series (Multivariate LSTM)

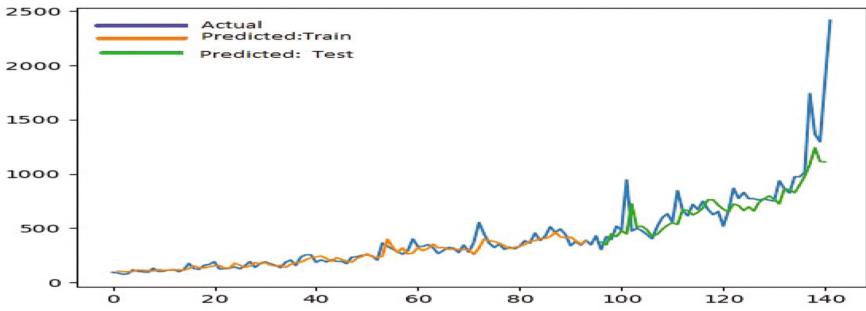


Plots of Predicted Vs. Actual Bug Number Series (Multivariate LSTM+ Covariance)

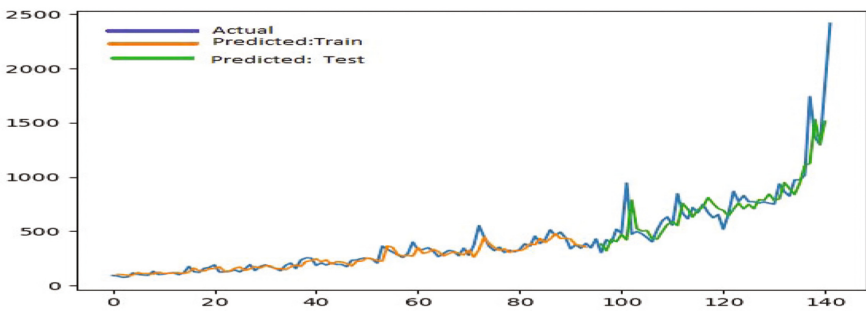
Fig. 7. Plots of predicted vs. actual bug number series: Mozilla



Plots of Predicted Vs. Actual Bug Number Series (Univariate LSTM)



Plots of Predicted Vs. Actual Bug Number Series (Multivariate LSTM)



Plots of Predicted Vs. Actual Bug Number Series (Multivariate LSTM+ Covariance)

Fig. 8. Plots of predicted vs. actual bug number series: Debian

5 Conclusion and Future Work

In this paper, we have used LSTM univariate and multivariate modelling for bug number prediction for three different open source software applications. We observed that multivariate LSTM gives more accurate results than univariate LSTM model. We also observe that after adding the covariance factor the accuracy further increases. Covariance can be considered as a goodness factor for the LSTM multivariate model for bug number prediction.

In future, we will apply advanced machine learning technique for bug number prediction. We will also apply prediction interval-based estimation approach for modelling the bug numbers. It will remove the chance of uncertainty associated with the point estimate and can be a more reliable model for bug number prediction.

References

1. Zimmermann, T., Nagappan, N., Zeller, A.: Predicting bugs from history. In: *Software Evolution*, pp. 69–88. Springer, Heidelberg (2008)
2. Herraiz, I., Gonzalez-Barahona, J.M., Robles, G.: Forecasting the number of changes in Eclipse using time series analysis. In: *Fourth International Workshop on ICSE Workshops on Mining Software Repositories, MSR 2007*, p. 32. IEEE, May 2007
3. Wu, W., Zhang, W., Yang, Y., Wang, Q.: Time series analysis for bug number prediction. In: *2010 2nd International Conference on Software Engineering and Data Mining (SEDM)*, pp. 589–596. IEEE, June 2010
4. Zhang, H.: An initial study of the growth of eclipse defects. In: *Proceedings of the 2008 International Working Conference on Mining Software Repositories*, pp. 141–144. ACM, May 2008
5. Pati, J., Shukla, K.K.: A nonlinear ARIMA technique for Debian bug number prediction. In: *Proceedings of the International Conference on Advances in Computer and Electronics Technology - ACET 2014, SEEK-DL 2014, September 2014*. doi:[10.15224/978-1-63248-024-8-14](https://doi.org/10.15224/978-1-63248-024-8-14)
6. Pati, J., Shukla, K.K.: A comparison of ARIMA, neural network and a hybrid technique for Debian bug number prediction. In: *2014 International Conference on Computer and Communication Technology (ICCCT)*, pp. 47–53. IEEE, September 2014
7. Debian Bug Tracking System. <http://www.debian.org/Bugs/>
8. Eclipse Bug Tracking System. <https://bugs.eclipse.org/bugs/>
9. Mozilla Bug Tracking System. <https://bugzilla.mozilla.org/>
10. Zhao, X., Wang, C., Yang, Z., Zhang, Y., Yuan, X.: Online news emotion prediction with bidirectional LSTM. In: *International Conference on Web-Age Information Management*, pp. 238–250. Springer International Publishing, June 2016
11. Karpathy, A., Johnson, J., Fei-Fei, L.: Visualizing and understanding recurrent networks. arXiv preprint [arXiv:1506.02078](https://arxiv.org/abs/1506.02078) (2015)
12. Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R., Schmidhuber, J.: LSTM: a search space odyssey. *IEEE Trans. Neural Netw. Learn. Syst.* (2016)
13. Gers, F.A., Schmidhuber, J., Cummins, F.: Learning to forget: continual prediction with LSTM. *Neural Comput.* **12**(10), 2451–2471 (2000)

14. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
15. Graves, A.: Supervised sequence labelling. In: *Supervised Sequence Labelling with Recurrent Neural Networks*, pp. 5–13. Springer, Heidelberg (2012)
16. Deep Learning 0.1 documentation. <http://deeplearning.net/tutorial/lstm.html>

Direct Demodulator for Amplitude Modulated Signals Using Artificial Neural Network

Vineetha K.V.^{1(✉)} and Dhanesh G. Kurup²

¹ Department of Computer Science and Engineering,
Amrita University, Bengaluru, India
jain.vineetha@blr.amrita.edu

² Department of Electronics and Communication Engineering,
Amrita University, Bengaluru, India

Abstract. This article, presents an Artificial Neural Network (ANN) based high speed demodulator, capable of demodulating amplitude modulated signals. The ANN is developed using Multilayer Perceptron (MLP) based Neural Network. We also introduce a pre-processing method for faster training of the ANN and a method training ANN using random modulating signal. Test results are presented for modulating signals such as triangular wave and square wave. We show that, the unique preprocessing technique introduced in this article enables us to achieve faster training of the ANN.

1 Introduction

In recent years, the development of communication technology [1] and the demand for very high data rate in the field of communications has been growing rapidly. In general, higher data rates means higher processing speed requirement for the system. However, the technologies such as wireless sensor network (WSN) and radio frequency identification (RFID) uses very small carrier frequency and very small data rates [2]. For example, there is a standard for RFID with carrier frequency of 13.56 MHz [3,4]. In such scenarios, direct digitization [5] of signals is possible and we will be able to process the digitized data in software or hardware. The key hardware component for digitization is the analog to digital converter (ADC) which converts time signals from continuous to discrete, binary coded format. Advances in high speed communications and software radio development have necessitated ADC performance improvement in sampling rates of the order of 10 Million samples per second (Ms/s) [6]–[8].

In a demodulation process, the amplitude of the modulated signal or phase of the modulated signal or both amplitude and phase of the modulated signal, is retrieved [9]. The resulting signal is further processed to derive the binary bits if the modulation is digital or analog signal if the modulation is analog modulation [10]. State of the art demodulators are all-digital, in the sense that, the demodulation is done after IF to digital conversion [11]. Therefore, demodulation can be executed in digital platforms such as digital signal processors (DSP),

field programmable gate arrays (FPGA) or application specific integrated circuits (ASIC) [12]. For instance, in [13], demodulation of higher order quadrature amplitude modulation (QAM) signal is carried out using a high data-rate parallel demodulator implementation on FPGA platform. The paper [14] discuss the multicomponent AM-FM demodulation techniques based on the Hilbert transform. This approach is all digital and does not need the complex filter optimizations and has higher performance.

The ANN [15] is an input-output mapping architecture with multiple layers of neurons and weighted interconnections for classification and regression of data. ANN offers many advantages such as high speed of operation, decreased delay, high reliability and resilience against noisy data. Therefore, ANN enables massive computations once trained to perform a certain action. Among the ANN topologies, the Multi-Layer Perceptron (MLP) has the simplest architecture with neurons associated with activation functions and threshold values. The Neurons in one layer MLP are connected to all other neurons in the following layer through the links, which represents the connection weights.

In [16], an MLP architecture is used to design a demodulator for telecommunication signals for Universal Mobile Telecommunications System (UMTS) Terrestrial Radio Access. In [17] the Frequency Shift Keying(FSK) signal is demodulated using designed ANN. The recurrent network, Elman Artificial Neural Network(EANN) having four-layer network is used here. The pattern recognition characteristics based ANN demodulator using Gaussian Minimum Shift-Keying(GMSK) signals is discussed in [18]. An ANN demodulator to demodulate binary frequency shift keying signal is discussed in [19]. The trained ANN can be further used to train any type of modulation methods with no change in the hardware. Compared with other ANN demodulators, we can train the ANN in [19] faster and with less training data, resulting in very low bit error rates (BER). The paper [20] discusses about a probabilistic neural network which can be trained for different digital modulation schemes to detect incoming data with no change in hardware. Therefore, the ANN in [20] can be used in the applications where fast training is required and with small Bit Error Rate (BER).

It is to be noted that, most of the ANN based demodulation methods described above requires Discrete Fourier Transform (DFT) and(or) Inverse Discrete Fourier Transform (IDFT) tools which requires large memory and buffering. In this article, we present a direct demodulation architecture based on MLP where radio frequency (RF) to intermediate frequency (IF) conversion is avoided. This enables us to convert the modulated carrier directly to baseband waveform. The training is carried out using the widely used back propagation algorithms using random generated input data. Although, we only study the ANN for an amplitude modulated signal, the method can be extended to digital IQ (In-phase and Quadrature demodulators) for demodulating the digitally modulated signals [21].

The proposed demodulator can efficiently process high frequency modulated carrier corrupted with noise. For implementing the proposed demodulator, we used the open source FANN library [22].

The rest of the paper is divided into the following subsections. Section 2 gives the mathematical formulations used for the training and the theoretical aspects. Section 3 illustrates test results for different types of input signals. Finally, Sect. 4 describes the conclusion.

2 Theory

The proposed ANN based demodulator consists of one input layer, one output layer and one or two hidden layers. The input of demodulator can be defined as,

$$Y(t) = I(t) \cos(2\pi f_c t) \quad (1)$$

where, $Y(t)$ is the modulated carrier signal, $I(t)$ is the in-phase component of a signal, f_c is the carrier frequency and t is the time. The goal of the demodulator is to extract $I(t)$ from $Y(t)$.

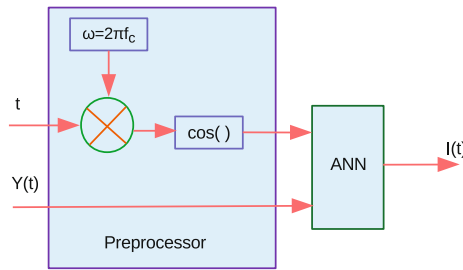


Fig. 1. Block diagram of proposed demodulator

Figure 1 shows the demodulator with embedded ANN. As we can see from Fig. 1, the demodulator consists of a preprocessor block and an ANN block. The preprocessor block has a multiplier in it, which is used to multiply the time t component with ω_c component to generate $\omega_c t$ component, where $\omega_c = 2\pi f_c$. After generating the $\omega_c t$, Cosine operator is applied further to $\omega_c t$, which will be the input to our ANN block. The preprocessing method used here is the computation of the cosine part of $\omega_c t$ component. The pre-processing method described enabled us to train the neural network in a very fast manner. One of the remarkable advantage of the proposed demodulator compared to previously published ANN is the, speed in training due to the preprocessing method introduced. The ANN block of Fig. 1 is an MLP which was trained using the Quick Propagation algorithm. The $Y(t)$ component and the preprocessed $\omega_c t$ are the two given inputs to the ANN block. The output of the ANN block will be the $I(t)$ component. We have tried both single as well as two hidden layers with the total number of neurons in the hidden layers from 10 to 50 neurons. For generating the training data, we used two uniform random number generators. The uniform random number generator helps us to generate the numbers in continuous manner within the given range. Initially we generate time t and In-phase component $I(t)$ using random number generator.

3 Results

For training the ANN with sigmoid activation function at neurons, the training algorithm used is Quick propagation with learning rate 0.1. The time range is chosen in the range $t = [0 : 4]$ s. The range of $I(t)$ component is set as -1.2 v to 1.2 v. This training input data is generated using two uniform random number generators. Using this data we calculate $Y(t)$ which is the input of the ANN as shown in Fig. 1. The actual $I(t)$ and generated ANN $I(t)$ are compared, both looks almost similar. Here we prove that the $I(t)$ which is extracted from $Y(t)$ through ANN and the actual $I(t)$ are same. Hence this proves that the training method what we used is working properly and this can be used further for processing various types of signals. The minimum time is set as zero and the maximum time is the number of cycles, which is defined as 4. Later cross-validation, we have used triangular wave or square wave as $I(t)$ component. We fix the range of $I(t)$ component between -1.2 v to 1.2 v.

Figure 2 shows the plot of number of Neurons versus Root Mean Square Error (RMSE). Here the trained data is validated using single hidden layer and two hidden layers. The RMSE is generated for both the single and two hidden layers, where neurons varying in the range of 10–50 neurons. Since, for number of neurons >20 , the performance improvement was minimal, we used 20 neurons for final ANN based demodulator.

The results shows that, two hidden layers will give the least error compared to the single hidden layer. The RMSE is 0.02 here. Hence the proposed training method of ANN can efficiently process high frequency modulated carrier signal with less error.

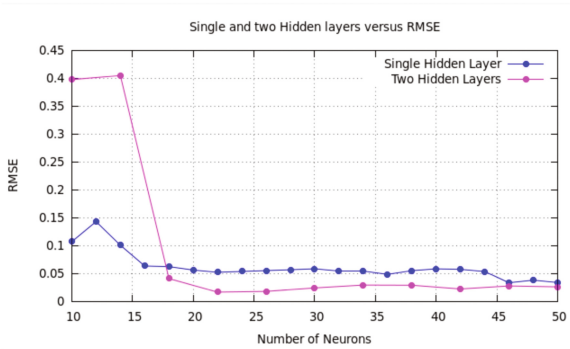


Fig. 2. Data validation using single hidden layer and two hidden layers

The trained network is tested for a known signal where the in-phase component of triangle wave and square wave are tested. The following plots shows the various results generated after the test.

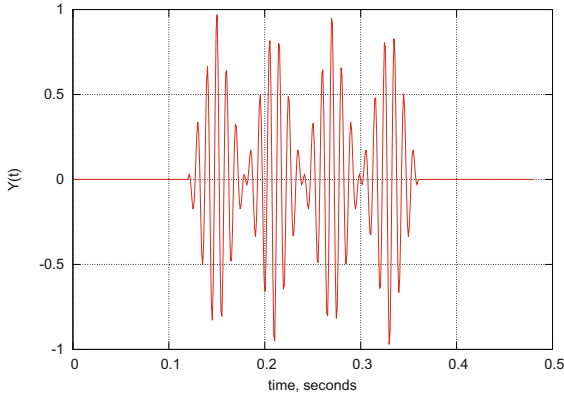


Fig. 3. Modulated carrier signal $Y(t)$

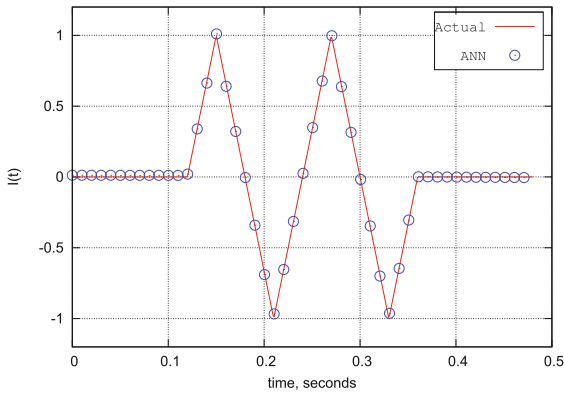


Fig. 4. Comparison of output of ANN (single hidden layer) and required output.

Figure 3 shows modulated carrier and Fig. 4 shows the comparison of output of ANN using single hidden layer and the required output, where the tested signal is a triangular wave. The plot is of time t versus $I(t)$ component. The triangular wave, that is the actual $I(t)$ component is compared with the ANN generated $I(t)$ component. Both the $I(t)$ components looks almost similar. Hence our result shows that the proposed demodulator based on ANN can efficiently process high frequency modulated carrier signal in a very fast manner with less error.

Figure 5 shows the comparison of output of ANN using two hidden layers and the required output, where the tested signal is a triangular wave. The plot is of time t versus $I(t)$ component. The triangular wave, that is the actual $I(t)$ component is compared with the ANN generated $I(t)$ component. Both the $I(t)$ components looks almost similar. Hence our result shows that the proposed demodulator based on ANN can efficiently process high frequency modulated carrier signal in a very fast manner with less error.

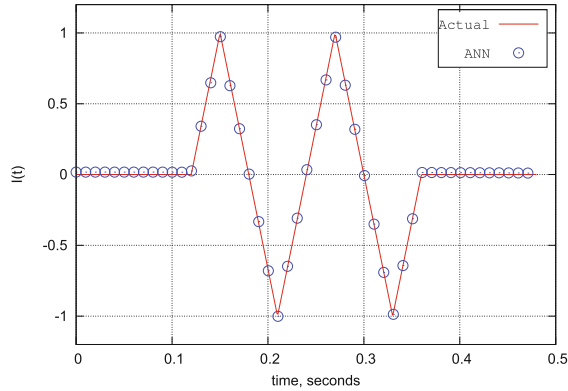


Fig. 5. Comparison of output of ANN (two hidden layers) and required output.

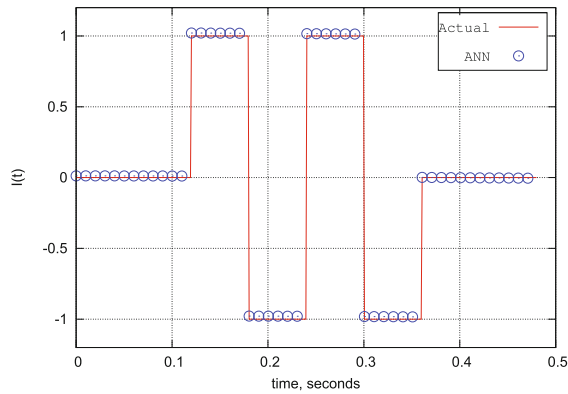


Fig. 6. Comparison of output of ANN (single hidden layer) and required output.

Figure 6 shows the comparison of output of ANN using single hidden layer and the required output, where the tested signal is a square wave. The plot is of time t versus $I(t)$ component. The square wave, that is the actual $I(t)$ component is compared with the ANN generated $I(t)$ component. Both the $I(t)$ components looks almost similar. Hence our result shows that the proposed demodulator based on ANN can efficiently process high frequency modulated carrier signal in a very fast manner with less error.

The Fig. 7 shows the comparison of output of ANN using two hidden layers and the required output, where the tested signal is a square wave. The plot is of time t versus $I(t)$ component. The square wave, that is the actual $I(t)$ component is compared with the ANN generated $I(t)$ component. Both the $I(t)$ components looks almost similar. Hence our result shows that the proposed demodulator based on ANN can efficiently process high frequency modulated carrier signal in a very fast manner with less error rate.

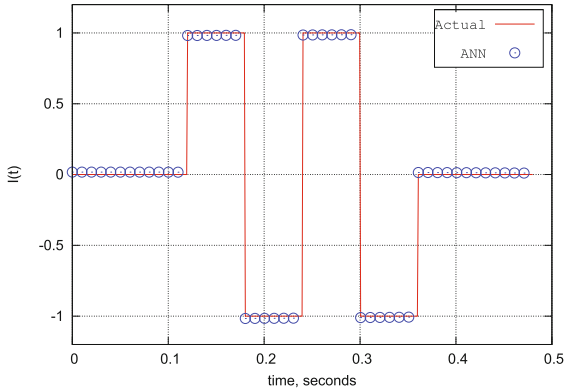


Fig. 7. Comparison of output of ANN (two hidden layers) and required output.

4 Conclusion

The unique preprocessing feature of ANN for realizing demodulator, introduced in this paper, enables us to achieve faster training and high speed of operation. The results shows that direct demodulation of an amplitude modulated signal based on MLP enables us to convert the modulated carrier directly to baseband waveform, by avoiding the conversion from RF to IF. We can easily convert the proposed modulator to digital IQ demodulators. The proposed demodulator can also be implemented in digital platforms such as DSP and FPGA.

References

1. Verhelst, M., Bahai, A.: Where analog meets digital, analog-to-information conversion and beyond. *IEEE Solid-State Circ. Mag.* **7**(3), 67–80 (2015)
2. Kaur, M., Sandhu, M., Mohan, N., Sandhu, P.S.: RFID technology principles, advantages, limitations and its applications. *Int. J. Comput. Elect. Eng.* **3**, 1793–8163 (2011)
3. Finkenzeller, K.: *RFID Handbook - Fundamentals and Applications in Contactless Smart Cards, Radio Frequency Identification and Near-Field Communication*, 3rd edn. Wiley, Hoboken (2010)
4. Reshmi, K., Dhanesh, G.K.: Implementation aspects of a new RFID anti-collision algorithm. In: *Conference Proceedings TENCON-2016, Singapore*, pp. 127–129, December 2016
5. Khan, S.: Digitization and its impact on economy. *Int. J. Digital Libr. Serv.* **5**(2) (2015)
6. Mitola, J.: The software radio architecture. *IEEE Commun. Mag.* **33**, 26–38 (1995)
7. Wepman, J.A.: Analog-to-digital converters and their applications in radio receivers. *IEEE Commun. Mag.* **33**, 39–45 (1995)
8. Baines, R.: The DSP bottleneck. *IEEE Commun. Mag.* **33**, 46–54 (1995)

9. Singh, A., Kumar, D.S., Venkateswaran, G., Manjukrishna, S., Singh, A.K., Kurup, D.G.: Design and experimental characterization of a bandpass sampling receiver. In: International Conference on Communication Systems, ICCS-2015. American Institute of Physics (AIP), Pilani (2015)
10. Adeleke, O.A., Abolade, R.O.: Modulation methods employed in digital communication: an analysis. *Int. J. Elect. Comput. Sci. IJECS-IJENS* **12**(3)
11. Walden, R.H.: Analog-to-digital converter survey and analysis. *IEEE J. Sel. Areas Commun.* **17**(4) (1999)
12. Sklyarov, V., Skliarova, I., Sudnitson, A.: FPGA-based systems in information and communication. In: 5th International Conference on Application of Information and Communication Technologies (AICT), pp. 1–5 (2011)
13. Lin, C., Shaol, B., Zhang, J.: A high data rate parallel demodulator suited to FPGA implementation. In: International Symposium on Intelligent Signal Processing and Communication Systems, pp. 6–8, December 2010
14. Gianfeli, F., Turchetti, C., Crippa, P.: Multicomponent AM-FM demodulation: the state of the art after the development of the iterated hilbert transform. In: IEEE International Conference on Signal Processing and Communications, ICSPC 2007, pp. 24–27, November 2007
15. Uhrig, R.E.: Introduction to artificial neural networks. In: 21st International Conference on Industrial Electronics, Control, and Instrumentation, Proceedings of the IEEE IECON, vol. 1, pp. 33–37, November 1995
16. Grimaldi, D.: ANN based demodulator for UMTS signal measurements. *Measur. J.* **39**(10), 877–883 (2006)
17. Li, M., Zhong, H., Li, M.: Neural network demodulator for frequency shift keying. In: International Conference on Computer Science and Soft Engineering, vol. 4, pp. 843–846 (2008)
18. Aiello, A., Grimaldi, D., Rapuano, S.: GMSK neural network based demodulator. In: International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, Foros, Ukraine (2001)
19. Amini, M.R., Moghadasi, M., Fatehi, I.: A BFSK neural network demodulator with fasttraining hints. In: Second International Conference on Communication Software and Networks (2010)
20. Amini, M.R., Balarastaghi, E., Branch, B.: Universal neural network demodulator for software defined radio. *ACSIT Int. J. Eng. Technol.* **3**(3) (2011)
21. Mutha, S., Roblin, P., Chaillot, D., Yang, X., Kim, J., Strahler, J., Rojas, R., Volakis, J.: Technique for joint balancing of IQ modulator-demodulator chains in wireless transmitters. In: IEEE MTT-S International Microwave Symposium Digest, pp. 221–224 (2009)
22. <http://www.leenissen.dk/fann/html/files>

Real-Time Detection of Atrial Fibrillation from Short Time Single Lead ECG Traces Using Recurrent Neural Networks

V.G. Sujadevi^(✉), K.P. Soman, and R. Vinayakumar

Centre for Computational Engineering and Networking (CEN),
Amrita School of Engineering, Amrita Vishwa Vidyapeetham,
Amrita University, Coimbatore, India
sujapraba@gmail.com, kp_soman@amrita.edu, vinayakumarr77@gmail.com

Abstract. Atrial fibrillation (AF) is the predominant type of cardiac arrhythmia affecting more than 45 Million individuals globally. It is one of the leading contributors of strokes and hence detecting them in real-time is of paramount importance for early intervention. Traditional methods require long ECG traces and tedious preprocessing for accurate diagnosis. In this paper, we explore and employ deep learning methods such as RNN, LSTM and GRU to detect the Atrial Fibrillation (AF) faster in the given electrocardiogram traces. For this study, we used one of the well-known publicly available MIT-BIH Physionet dataset. To the best of our knowledge this is the first time Deep learning has been employed to detect the Atrial Fibrillation in real-time. Based on our experiments RNN, LSTM and GRU offer the accuracy of 0.950, 1.000 and 1.000 respectively. Our methodology does not require any de-noising, other filtering and preprocessing methods. Results are encouraging enough to begin clinical trials for the real-time detection of AF that will be highly beneficial in the scenarios of ambulatory, intensive care units and for real-time detection of AF for life saving implantable defibrillators.

1 Introduction

Atrial fibrillation (AF) is a disorder of the functioning of the heart's electrical system that is characterized by the irregular beating of the heart [1]. Globally AF affects more than 35 Million people and the results of AF ranges from simple giddiness to mortality [1]. AF can be caused by several alterable and Non-alterable conditions. For example, valvular heart diseases and high blood pressure are shown to lead one to developing atrial fibrillation [2]. Other non-alterable conditions such as congenital heart disease and coronary artery diseases have been show to induce AF [2,3]. While several studies have been done for diagnosing AF and the treatment options, still several challenges have not been overcome [4]. One of the key factor in treating a subject for AF is to detect the presence of AF early enough primarily by ECG with the focus on prevention of the stroke which

is one of the deadly aftereffect of AD [5]. One of the most common techniques is to identify AF the absence of “P” wave in the 12 lead ECG [6].

Several methodologies and algorithms have been studied to identify the pathological ECG from the normal sinus rhythm [6,9]. Singular Value Decomposition (SVD) of wavelet coefficients and SVM has been employed to detect the arrhythmia with classification by Support Vector Machine, Linear Discriminant Analysis and Classification tree algorithms and found to have given good accuracy [7]. While it is important to detect the cardiac arrhythmias such as AF early, it is equally important to reduce the false positive detection, especially in intensive care unit clinical setting, where any alarm will be responded immediately [8]. To ensure the false alarm for the AF can be identified, several approaches including comparing with other parameter measurements such as arterial blood pressure pulse have been compared for false positive detection [8]. Several de-noising methods has been proposed with high accuracy with lower processing overhead [9,10]. Combined methods that can detect the QRS complex in noisy data and the accurate detection of P and T waves using sparsity filter and Gaussian derivative filter has been proposed with accuracy rate of 99.91% than existing methods [9]. After the recent surge in Deep learning based Neural networks, that gives higher accuracy and less processing overhead than the traditional methods, we investigate the efficacy of the several Deep learning techniques for discriminating the AF from Normal Sinus Rhythm (NSR). One of the motivations for us to explore Deep learning is to avoid the cumbersome pre-processing of the data such as de-noising of ECG signal [10] and filtering thereby aiming to reduce the complexity and computational requirements thus enabling real-time detection. The real-time detection has several applications in the health care and remote monitoring of the cardiac patients that suffer from arrhythmias such as AF.

The rest of sections of this paper are structured as follows. Section 2 discusses concepts of deep learning algorithms specifically RNN, LTM and GRU, Sect. 3 displays the proposed deep learning architecture, Sect. 4 discusses hyper parameter tuning and description of data base. The detailed evaluation results of all deep networks is displayed in Sect. 5. Section 6 discusses the future work and discussions. At last, conclusion is placed in Sect. 7.

2 Background

This section provides an intuitive understanding of recurrent neural network (RNN) and long short-term memory (LSTM) mathematically and followed by training mechanism of both RNN and LSTM networks.

2.1 Recurrent Neural Network (RNN)

Recurrent neural network (RNN) was introduced in initial time for time-series data modeling [11]. They are same as feed-forward networks (FFN) with an additional cyclic loop, as shown in Fig. 1. This cyclic loop carries out information

from one time-step to another. As a result, RNN are able to learn the temporal patterns, value at current time-step is estimated based on the past and present states. In general, RNN accepts $x = (x_1, x_2, \dots, x_T)$ (where $x_t \in R^d$ as input and maps to hidden input sequence $hd = (hd_1, hd_2, \dots, hd_T)$ and output sequence $op = (op_1, op_2, \dots, op_T)$ from $t = 1$ to T by iterating the following recursive equations.

$$h_t = G(w_{xhd}x_t + w_{hdhd}hd_{t-1} + b_{hd}) \tag{1}$$

$$op_t = sf(w_{hdop}hd_t + b_{op}) \tag{2}$$

Where w represents weight matrices, b represents bias vectors, G is an element wise non-linear activation function, specifically *sigmoid* and sf is an element wise non-linear activation function, specifically *sigmoid*.

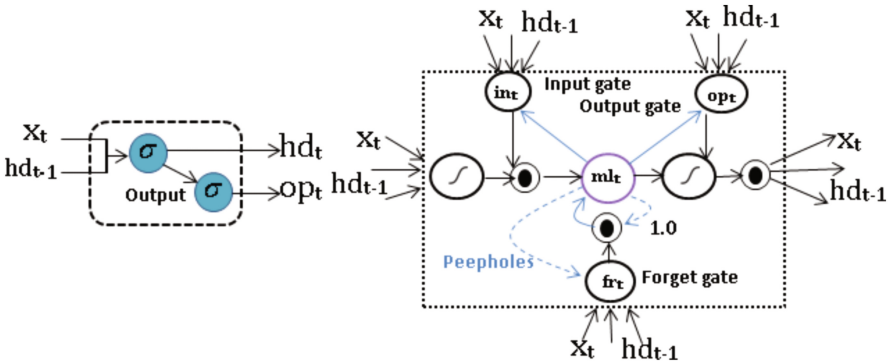


Fig. 1. Architecture of RNN unit (left) and LSTM memory block (right)

2.2 Training in RNN

Training RNN requires the network to be transformed to feed forward networks (FFN) using unfolding or unrolling. The unfolded RNN’s can be interpreted as the deep FFN’s without cyclic connections. Deep FFN’s usually consist of k hidden layers with the input sequence of length k , as shown in Fig. 2. The newly formed FFN’s are flexible for back-propagation. However, the network parameters in RNN are shared across all time-steps. Hence, to estimate the gradient at a particular time-step, rely on the current and as well as previous time-steps. This strategy is called as back-propagation through time (BPTT). However, it led to vanishing and exploding gradient problem while training RNN to learn long-term temporal dynamics of sequences of arbitrary length across many time-steps [12]. This is fundamentally due to the fact that gradient vector can grow or decay exponentially when propagating through many layers of RNN to learn long-term dependencies in time-steps. Further research brought out many variants to the RNN. In that, LSTM [12] emerged as a successful paradigm to handle long-term dependencies of sequences of arbitrary length.

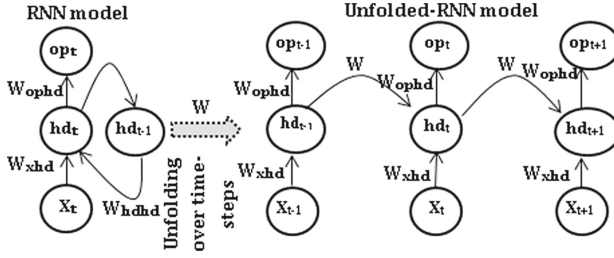


Fig. 2. RNN model and the unfolded-RNN model across time in forward direction

2.3 Long Short-Term Memory (LSTM)

Long short-term memory (LSTM) is an improved method of traditional RNN network that solves the vanishing and exploding gradient issue by forcing the constant error flow. LSTM has introduced a memory block instead of a simple RNN unit. A memory block is a subnet of LSTM architecture that contains one or more memory cell with a pair of adaptive multiplicative gates as input and output gate, as shown in Fig. 1. A memory block houses an information and updates them across time-steps based on the input and output gates. Input and output gate controls the input and output flow of information to a memory cell. Additionally, it is has a built-in value 1 for constant Error carousel (CEC). This value will be activated when in the absence of value from the outside signal. Most importantly, LSTM has performed well in learning long-range temporal dependencies in various long-standing artificial intelligence (AI) tasks [13]. As further studies on LSTM, additional components have been added to the existing LSTM architecture. [14] found that the internal values of a memory cell could increase without any limitations and to control them, they replaced CEC with the forget gate. Forget gate facilitates to forget the past value at a specific time step. Moreover, to learn the precise timing of the output, peephole connections are added from a memory cell to all of its adaptive multiplicative gates [15].

In general, LSTM accepts an input $x = (x_1, x_2, \dots, x_T)$, estimates output sequence by continuously updating the values of adaptive multiplicative units such as input (*in*), output (*op*) and forget gate (*fr*) on a memory cell (*ml*) in an iterate manner from $t = 1$ to T in the recurrent hidden layer of LSTM architecture. At each time-step T the LSTM recurrent hidden layer function is mathematically formulated as follows:

$$x_t, hd_{t-1}, ml_{t-1} \rightarrow hd_t, ml_t$$

$$in_t = \sigma(w_{xin}x_t + w_{hdin}hd_{t-1} + w_{mlin}ml_{t-1} + b_{in}) \tag{3}$$

$$fr_t = \sigma(w_{xfr}x_t + w_{hdfr}hd_{t-1} + w_{mlfr}ml_{t-1} + b_{fr}) \tag{4}$$

$$ml_t = fr_t \odot ml_{t-1} + i_t \odot \tanh(w_{xmi}x_t + w_{hdm}hd_{t-1} + b_{ml}) \tag{5}$$

$$op_t = SG(w_{xop}x_t + w_{hdop}hd_{t-1} + w_{mlop}ml_t + b_{op}) \tag{6}$$

$$hd_t = op_t \odot \tanh(ml_t) \tag{7}$$

where *in*, *op*, *fr*, *ml* term represents the input gate output gate, forget gate and a memory cell respectively. From the Fig. 1, we can say an LSTM network is composed of many components. Thus it ends up in more training cost. This might be one of the reasons to further enhancement of LSTM network.

2.4 Gated Recurrent Unit (GRU)

Gated recurrent unit (GRU) (see Fig. 3) is an improved network of LSTM [16]. It is simpler than LSTM and includes less computation. The computational flow of GRU is given below

$$x_t, hd_{t-1} \rightarrow hd_t$$

$$i_fr_t = \sigma(w_{xi_fr}x_t + w_{hdi_fr}hd_{t-1} + b_{i_fr})(Updategate) \tag{8}$$

$$fr_t = \sigma(w_{xfr}x_t + w_{hdf}hd_{t-1} + b_{fr})(Forgetorresetgate) \tag{9}$$

$$ml_t = \tanh(w_{xmi}x_t + w_{hdm}(fr \odot hd_{t-1}) + b_{ml})(Currentmemory) \tag{10}$$

$$hd_t = fr \odot hd_{t-1} + (1 - fr) \odot ml(Updatedmemory) \tag{11}$$

Formulae shows, unlike LSTM memory cell with a list of gates (input, output and forget), GRU only consist of gates (update and forget) that are collectively involve in balancing the interior flow of information of the units. Input gate (*in*) and forget gate (*fr*) are combined and formed a new gating unit typically called as update gate (*i_fr*). The update gate is mainly focus on to balance the state between the previous activation *ml* and the candidate activation *hd* without peephole connections and output activations. The forget gate resets the previous state *ml*.

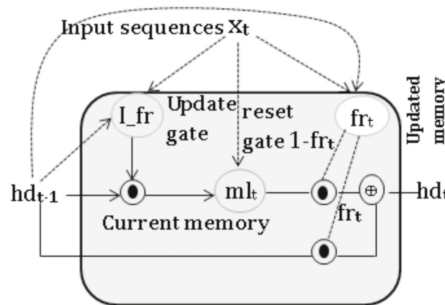


Fig. 3. Units in GRU

3 Network Architecture

The architecture for distinguishing a signal as normal sinus rhythm (NSR) and atrial fibrillation (AF) is displayed in Fig. 4. Unlike classical machine learning classifiers, the proposed system does not rely on any of feature engineering mechanism. Instead, it accepts raw input signal as such and that will be fed to recurrent layers such as RNN, LSTM and GRU to obtain optimal feature representation including long-term dependencies. The newly formed feature representation is passed to dense layer for classifying a signal as NSR and AF using sigmoid non-linear activation function.

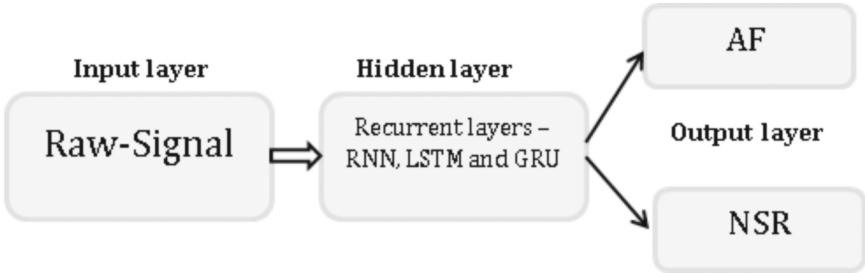


Fig. 4. Architecture of proposed system for normal sinus rhythm and atrial fibrillation

4 Experiments

All trails of experiments are run on Graphics processing unit (GPU) enabled TensorFlow (r0.11.0) [17] computational framework in single NVidia GK110BGL Tesla k40 in Ubuntu 14.04 operating system (OS). LSTM network consist of a set of parameters such as learning-rate, memory blocks, number of hidden layers, the number of epochs etc. To choose a good parameter value, the various configurations of network parameters of LSTM are used in each experiment. BPTT technique is used in training the LSTM model and the memory cells of LSTM used *tanh* as input and output squashing function, *sigmoid* for gates.

4.1 Description of Dataset

We used the publically available raw signals of Atrial fibrillation (AF) and normal sinus rhythm (NSR) from MITBIH Physionet; MIT-BIH Atrial Fibrillation Database and MIT-BIH Normal Sinus Rhythm Database [21]. Each trace was 60 s long and was sampled at 250 Hz for AF and 128 Hz for NSR. The signals were not pre-processed for noise-removal etc. A single lead ECG wave form of NSR and AF of MITBIH Physionet data base is depicted in Fig. 5(a) and (b) respectively. The detailed statistics of the MITBIH Physionet database is displayed in Table 1.

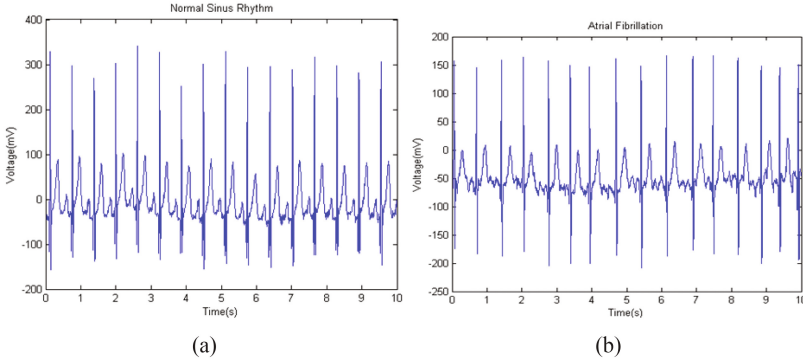


Fig. 5. (a) A single lead ECG wave form of normal sinus rhythm, (b) A single lead ECG wave form with atrial fibrillation

Table 1. Statistics of MIT-BIH Atrial Fibrillation Database and MIT-BIH Normal Sinus Rhythm Database

Name of signal	Number of signals
AF	25
NSR	25

4.2 Hyper Parameter Tuning in LSTM Network

To identify suitable network parameters for LSTM, initially we started to experiment with a moderately-sized LSTM network. A moderately-sized LSTM network contains a recurrent hidden layer and dense layer with sigmoid non-linear activation function. A recurrent hidden layer has 2 memory cells containing one memory cell each and fixed learning rate 0.1. 2 trials of experiments in 5-fold cross validation are conducted for each parameter of memory blocks varying in the range [2–64]. Each trails of experiment are run up to 100 epochs. 64 memory blocks in recurrent hidden LSTM layer showed better accuracy in 5-fold cross-validation configuration setting in comparison to the other parameters of memory blocks. The same experiments are followed for RNN and GRU. Among all LSTM and GRU has showed highest accuracy but RNN performance was comparable to them. GRU performance was good in comparison to LSTM in terms of training cost.

5 Evaluation Results

Based on the obtained results during hyper parameter tuning in previous section, we fixed the number of blocks to 64 in recurrent hidden LSTM layer and recurrent hidden GRU layer, 64 units in recurrent hidden RNN layer, learning rate 0.1. Using these deep networks, experiments are conducted on MIT-BIH test dataset.

The results has outperformed the previously stated results, more importantly recent one [19]. The detailed test results as displayed in Table 2.

Table 2. Summary of test results

Algorithm	Accuracy	Precision	Recall	F-score
RNN	0.95	1.00	0.889	0.941
LSTM	1.00	1.00	1.00	1.00
GRU	1.00	1.000	1.00	1.00

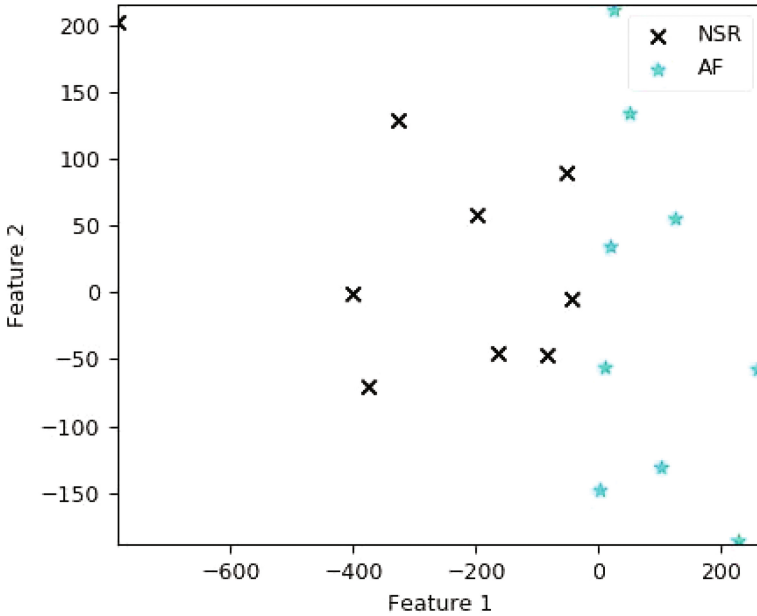


Fig. 6. 10 samples of each classes of NSR and AF with their corresponding activation values of the last hidden layer neurons are represented using 2-dimensional linear projection (PCA). Note that the samples are clustered based on the similarity in activation values

Generally, a deep network passes given raw signals to various deep layers. The non-linear activation function in each layer enables to discriminate the signal as either AF or NSR. The last layer activation values should maximally separate various classes by using the learned feature representations. To visualize, the high-dimensional neuron unit activation vectors of last layer i.e. before the sigmoid activation function layer were redirected to t-SNE [20]. t-SNE is a dimensionality reduction mechanism that reduced the high-dimensional hidden layer feature representation into two-dimensional representation. The 2D vectors are finally plotted using Scikit-learn, as shown in Fig.6. In Fig.6 ECG traces

with similar characteristics have clustered together. More importantly, the ECG traces of both AF and NSR have appeared completely in separate clusters. This infers that the LSTM network has learnt well.

6 Future Work and Discussions

In this work, the effectiveness of deep learning approaches such as RNN, LSTM and GRU are discussed for classifying a signal as either Atrial Fibrillation (AF) or normal sinus rhythm (NSR). The outcome of the proposed method is robust. The existing studies on these datasets have empirically relied on various feature engineering mechanisms such as P-wave analysis including R-wave detection as initial mechanism in addition to noise filtering approach as preprocessing step to make ECG for constructive rhythm analysis. The significant advantage of the proposed method is not to rely on any feature engineering and noise filtering approaches. Based on results, we claim that our method outperforms the other published methods in effectively classifying a signal as AF or NSR. Though the deep network methods showed significant results, we lack in showing the inner mechanics of the deep models. This can be achieved by transforming the non-linearity to linearized form, thereby computing the Eigen values and Eigen vectors on them across time-steps [18]. As part of our future work we will employ the same methodologies to our real world dataset that has been collected from hospital that specializes in Cardiac care. Moreover, the computational performance of each deep networks will be discussed.

7 Conclusion


This paper has presented a novel deep learning based mechanism such as RNN, LSTM and GRU that robustly distinguished AF and NSR on a single lead ECG. All the deep learning methods have performed well, mostly LSTM and GRU outperformed RNN and GRU takes less training cost in comparison to LSTM. The proposed method is considered as more accurate in real-time ECG classification because it doesn't rely on any feature engineering mechanisms.

References

1. Go, A.S., Hylek, E.M., Phillips, K.A., Chang, Y., Henault, L.E., Selby, J.V., Singer, D.E.: Prevalence of diagnosed atrial fibrillation in adults: national implications for rhythm management and stroke prevention: the AnTicoagulation and Risk Factors in Atrial Fibrillation (ATRIA) study. *JAMA* **285**(18), 2370–2375 (2001)
2. Anumonwo, J.M., Kalifa, J.: Risk factors and genetics of atrial fibrillation. *Cardiol. Clin.* **32**(4), 485–494 (2014)
3. Nguyen, T.N., Hilmer, S.N., Cumming, R.G.: Review of epidemiology and management of atrial fibrillation in developing countries. *Int. J. Cardiol.* **167**(6), 2412–2420 (2013)

4. Calkins, H., Kuck, K.H., Cappato, R., Brugada, J., Camm, A.J., et al.: 2012 HRS/EHRA/ECAS Expert Consensus Statement on Catheter and Surgical Ablation of Atrial Fibrillation. *Heart Rhythm* **9**, 632–696.e621 (2012)
5. Ferguson, C., Inglis, S.C., Newton, P.J., Middleton, S., Macdonald, P.S., Davidson, P.M.: Atrial fibrillation: stroke prevention in focus. *ACC* **27**(2), 92–98 (2013)
6. McManus, D.D., Lee, J., Maitas, O., Esa, N., Pidikiti, R., Carlucci, A., Harrington, J., Mick, E., Chon, K.H.: A novel application for the detection of an irregular pulse using an iPhone 4S in patients with atrial fibrillation. *Heart Rhythm* **10**(3), 315–319 (2013)
7. Peterek, T., Zaorálek, L., Dohnálek, P., Gajdos, P.: Recognition of pathological beats in ECG signals based on singular value decomposition of wavelet coefficients and support vector machine. In: 2015 38th International Conference on Telecommunications and Signal Processing (TSP), Prague, pp. 1–5 (2015)
8. Couto, P., Ramalho, R., Rodrigues, R.: Suppression of false arrhythmia alarms using ECG and pulsatile waveforms. In: Computing in Cardiology Conference (CinC), Nice, pp. 749–752 (2015)
9. Manikandan, M.S., Ramkumar, B.: Straightforward and robust QRS detection algorithm for wearable cardiac monitor. *Healthc. Technol. Lett.* **1**(1), 40–44 (2014)
10. Mohan, N., Sachin Kumar, S., Poornachandran, P., Soman, K.P.: Modified variational mode decomposition for power line interference removal in ECG signals. *Int. J. Electr. Comput. Eng.* **6**, 151–159 (2016)
11. Elman, J.L.: Finding structure in time. *Cogn. Sci.* **14**(2), 179–211 (1990)
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
13. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**(7553), 436–444 (2015)
14. Gers, F.A., Schmidhuber, J., Cummins, F.: Learning to forget: continual prediction with LSTM. *Neural Comput.* **12**(10), 2451–2471 (2000)
15. Gers, F.A., Schraudolph, N.N., Schmidhuber, J.: Learning precise timing with LSTM recurrent networks. *J. Mach. Learn. Res.* **3**(1), 115–143 (2003)
16. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint [arXiv:1412.3555](https://arxiv.org/abs/1412.3555) (2014)
17. Abadi, M., et al.: TensorFlow: a system for large-scale machine learning. In: Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI), Savannah, Georgia, USA (2016)
18. Moazzezi, R.: Change-based population coding. Ph.D. thesis, UCL (University College London) (2011)
19. Arunachalam, S.P., Annoni, E.M., Kapa, S., Mulpuru, S.K., Friedman, P.A., Tolkacheva, E.G.: Multiscale frequency technique robustly discriminates normal sinus rhythm and atrial fibrillation. <https://www.researchgate.net/publication/316912116>
20. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**(Nov), 2579–2605 (2008)
21. Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.-K., Stanley, H.E.: PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* **101**(23), e215–e220 (2000)

StyloLIT: Stylometry and Location Indicative Terms Based Geographic Location Estimation Using Convolutional Neural Networks

K. Surendran, O.P. Harilal, P. Hrudyap, and Poornachandaran Prabakaran 

Amrita Center for Cybersecurity Systems and Networks, Amrita School of Engineering,
Amritapuri, Amrita Vishwa Vidyapeetham, Amrita University, Amritapuri, India
{surendrank, harilalop, hrudyap, praba}@am.amrita.edu

Abstract. Estimation of geographic location information of users from social media portals such as twitter plays a vital role in areas such as disaster management, marketing, cyber forensics etc. At the same time increasing data privacy concerns forced the social media sites to make the sharing of geographic location as the opt-in feature, also increasing user awareness about privacy prevents the users from disclosing their location details. However, most users leave footprints unknowingly that could be used to identify their approximate location information. Since it is observed that social media users from multiple locations possess diversity in their expression of language, we propose a two level approach involving stylometry and location indicative terms to address this problem. Experimental results shows that our approach outperforms the current state of the art in predicting the geographical location of twitter users purely based on their text content.

1 Introduction

Social media sites such as Twitter and Facebook have become predominant tools for online users to share content that ranges from simple text to rich media. The growth of the user generated data is exponential [1]. For example, on an average, twitter receives around 6000 posts per second that corresponds to 500 million posts per day. In addition to connecting users, Social networking site also serves as news outlets replacing traditional media [2, 3]. Most of the social media services allow its users to add their geographic location information [4, 5] by utilizing the GPS of the smartphone with the consent from user [6] or manually, location of user's choice is selected. It provides an opportunity for studying the geospatial imprints left over by millions of people in their social media conversations. But, the provision for manual tagging of geographic location information create disambiguation as users are allowed to geo tag a post with a location of user's interest rather than their current location. Also, many of the social media conversations come with no geo-tag. It might lead to incorrect or less accurate outcomes. One way of overcoming this is to employ a geographic location prediction prior to the detailed analysis. A user's tweeting behaviour affect his geolocatability based on all the

features listed together in this research paper. From the perspective of privacy protection when a user is not using a popular term specific to that location and when user's writing style is different than the writing style that is specific to that location then that user is less prone to accurately estimating his Geolocation.

This paper has been organized into the following sections. In Sect. 2, set of related works are listed and described what is new and unique than previous works. In Sect. 3, the dataset and different metrics used for analysis are described. Methods employed are explained in detail in Sect. 4. Experimental results explained in Sects. 5 and 6 concludes the current research work with few suggestions for future enhancement.

2 Literature Survey

Our research work makes use of Convolutional Neural Networks (CNN) where its usage in data classification has been recognised for its unsupervised learning with effective results, Anil et al. [26] and Athira et al. [27] works shows that how CNN plays important role in data classification effectively. Estimating geographic location information of users from social media platforms has been an extensive research area in the recent years. There have been different types of approaches followed by researchers for addressing this problem, which include content based approach to analysing user's social network. This research work is purely based on text data which is collected from social media. Cheng et al. [8] proposes a lattice-based neighbourhood smoothing model for refining a user's location estimation after identifying words in tweets with a strong local geo-scope using a classification approach. A frequency based probabilistic approach followed by Chandra et al. [9], considered the terms in reply-tweet to the recipient of the reply-tweet rather than to the user who posted the reply-tweet message. Bo et al. [10], employ a method for automatically learning the location indicative words via feature selection methods to get better accuracy in geographical location estimation. Hecht et al.'s research work [21] shows that stylometry based classifier with embedded knowledge-base can give better accuracy. Sakaki et al. [25] has proposed a system which makes use of keywords from tweets for real-time event prediction. Miura et al. [28] work shows how a simple CNN can be employed to predict users/tweets geographic location. Our work not only considers the single words as location indicators but phrases are also taken into consideration which further enhances the geographic location estimation.

Several works [11–14] rely on the information on networks and communities for predicting the geographic location information of social media users. That is, they employ an approach which makes use of the connections among users. But, in recent years the social media users have become more privacy conscious on what they share and with whom they share by limiting their profile access as shown in Madden et al.'s work [15], Molok et al. [16] reveals the reason behind this which is because of increase in data leakage incidents. Thus, extracting the connection information become tedious task, this leads to limitations or lack of traces in estimating the geographic location of an author by solely through connection information. This is one of the motivation for us to look at the approaches such as using Stylometry and Location indicative terms.

For estimating the geographic location of social media user, is achieved by employing an approach which purely makes use of content based metrics. It has been observed that each geographical area have some phrases or words which are used only in those and nearby locations. Such words are termed as location indicative terms. They can be either words or terms. For example, for referring the night before Halloween, the phrase *mischief night* has been used more by people around the area of New York and nearby areas compared to others. For addressing group of two or more people, phrase *yous and youse* are used more by people from north east part of the United States [7]. These two metrics are used along with different stylometric features to estimate the user's geographic location from social media conversations. It is observed that this approach yields more accuracy compared to the existing state of the art technologies.

3 Dataset and Metrics

For this study our dataset consists of geo-tagged tweets across North America. Twitter streaming API is used to collect geo-tagged tweets for a given boundary of coordinates for acquiring the data. The collected data comprises of 145 million tweets across 3400 locations. From the collected data a language filter [17] is applied to remove the non-English contents and remove duplicate tweets. Since our dataset have predominant language as English, analysis of tweets are restricted to the same. Also, most of the features are specifically for English language. Further refinement process is carried out on the dataset by removing tweets that contain only non-informative phrases like "Hi", "How are you", hyperlinks etc. It is observed that many of the verified Twitter accounts are managed by group of people rather than individuals managing their unique accounts. Hence tweets from group accounts are excluded from the dataset. Then, tweets were evenly distributed across each location to avoid the bias towards a particular set of locations. The final curated dataset contains the 100 Million tweets from 2600 locations which is equivalent to approximately 39k tweets per location. The reason to restrict the dataset to North America is to avoid several dialects and languages [18] as feature extraction task is arduous when multiple languages and dialects are involved. Building a model using ground truth dataset [geo-tagged tweets] helps in better validation of the system than non-geo-tagged dataset as the stylometric features system learns will be specific to a location. Other than geolocation the tweet metadata provides place/city/state/country level details which can be used only to build dataset with different region level data like city wise/state wise/country wise dataset.

3.1 Stylometric Features

Stylometry is the study of linguistic and tonal style of texts. It's often used in authorship attribution on anonymous text contents [19]. Eisenstein et al.'s work [20] shows that how same topic is expressed differently by authors across different geographic locations. Hecht et al. [21] has crafted an algorithm named "CALGARI" which can classify users to their geographic location. These works shows that geographically author's style varies on same topic. But, it's obvious that any location have a number of social media users

who is having good command on languages. It is difficult to interpret the location specific writing style from the tweets of such users. Rangel et al. [22] tries to classify the authors based on writing style and map to different age ranges based on the same. Higher values in the age range implicates the good command on the language and it's decreasing when going to lower age ranges. This classification method is employed on our dataset and checked values of different metrics. It's observed that users in the lower age ranges possess more diversity in the writing style and in good number also. While on the higher age ranges similarity in the metric values across different locations has less information gain. This observed pattern is used to filter the tweets whose predicted age lies towards lower age ranges for stylometric based geographic location estimation. This is a dataset with size of 78 million tweets from 2100 locations.

Different stylometric based features are explained as follows. Apart from word and character based features, different syntactic, semantic and readability based metrics also taken into account. **Character level features** include average frequency of special characters like (~`@#\$\$%^&* _ - +=, . / \ ? <>) and different brackets ({}[]()) which are present in text. Word level features include average number of words per sentence, ratio of unique words and complex words used. Syntactic word features include frequencies of conjunction (but, because, etc.) and interjection (shh, phew, etc.) words. Also average count of coordinating (and, but, for, etc.) and subordinating (after, although, as, etc.) conjunction are calculated. Frequencies of adposition phrases including prepositional, postpositional and circum-positional phrases. And function words like article (a, an, etc.), pronoun (myself, she, etc.), auxiliary verb (can, may, etc.), particles(to fly, etc.), expletives(sentences starts with it, here and there) and pro-sentence(yes, no, okay, etc.) frequencies are calculated. **Semantic word features** like occurrence of soft words, greeting words, profanity are extracted. Emotional state features like positive, negative, etc. will help to classify mood of a person. Types of emotions which helps in classification author's mood are positive emotion (kind, hope, etc.), negative emotion (alone, afraid, etc.), tentative (guess, perhaps, etc.), negative (beg, abort etc.), positive (true, thank etc.), stopping (between, after etc.), agreeing (take, agree etc.), anxiety words (panic, worried, etc.). Frequency of occurrence of a proper sentence is considered for sentence level feature. **Readability metrics** [23] used for measuring the different parameters like easiness, complexity of text content. The readability measures that are considered in our research are Simple Measure of Gobbledygook (SMOG), Anderson's Readability Index (RIX), Automated Readability Index (ARI), Lycée International Xavier (LIX), Flesch Reading Ease (RE), Flesch Kincaid Grade Level (FKGL), Coleman Liau Index (CLI) and Gunning fog index (GOI) [23].

4 Methodology

In this section different methods used are explained which helps to estimate the geographical location of social media users purely based on content.

Algorithm.1. Extraction of location indicative phrases.

```

Set<dict_common phrases> Pdict_common
Initialize map<location, phrase_set> MLPs
Initialize map<phrase, location_set_with_weight> MPLs
For each location
  Initialize map<phrase, frequency> MPF
  For each post
    Extract word n-grams(p) with different length
    If p not in Pdict_common
      Update MPF.
  For each phrase(p) in MPF
    If p in MPLs
      If locations are nearby
        Update MPLs
      Else remove phrase from MPLs
    Else add to MPLs
  Update MLPs

```

Algorithm.2. Extraction of location indicative words.

```

Set<dict_common Words> Wdict_common
Initialize map<location, word_set> MLWs
Initialize map<word, location_set_with_weight> MWLs
For each location
  Initialize map<word, frequency> MWF
  For each post
    Extract each word(w)
    If w not in Wdict_common
      Update MWF
  For each word(w) in MWF
    If w in MWLs
      If locations are nearby
        Update MWLs
      Else remove word from MWLs
    Else add to MWLs
  Update MLWs

```

4.1 Extracting Location Indicative Terms

For achieving this, the dictionary and other common words are discarded first and all other terms mapped to the corresponding location after avoiding the overlapping of words across multiple locations. At the same time, the words to multiple locations are mapped in case those words are geographically close to each other. The algorithmic representation of the method is given in Algorithm 1. For extracting the phrases which are uniquely used in particular region or location, a word based n-gram approach is used. For each location, every tweet is iterated through to extract bi-grams, tri-grams, ..., n-grams and find common phrases used among different people. Then, as in words, discard overlapping of phrases across multiple locations while taking the common phrases

across neighbouring locations are taken into account. The algorithmic representation of the method is given in Algorithm 2. For finding the neighbours of a location, iteratively geographical distance to other locations in east, west, south and north directions are calculated and taking the locations with minimum distances as neighbours in each direction.

4.2 Stylometric Feature Extraction

The stylometric features from tweets of users of each location are extracted, using the formulae explained in the previous section. After normalization of the features, classification process is employed to build model.

4.3 Classification

Convolutional Neural Network (CNN) is used here for estimating geo-location from text. Convolutional neural network or ConvNet is feed-forward neural network with convolution layer and pooling layer.

Let $F_i \in R_n$ be the n-dimensional vector for the i-th feature in the input sentence. These features are character level, word level, sentence level and readability matrix which are V_1, V_2, V_3 and V_4 respectively. Each feature is generated from a series of analysis. Input for the next level can be calculated as, $F_i = f(w \cdot V_k + b)$ where $i = 1$ to k . Here k is the internal features of each vector V , w is the weight matrix and b is the bias. Over this feature map max-overtime pooling operation is applied in order to get the maximum feature value. This step will provide the relevant feature. These features are then passed to a fully connected softmax layer in order to get the output with probability corresponding to the detected class. Totally there are 52 features and 2600 output locations. Also an activation function \tanh [24] is used, which will help for faster convergence of training algorithm.

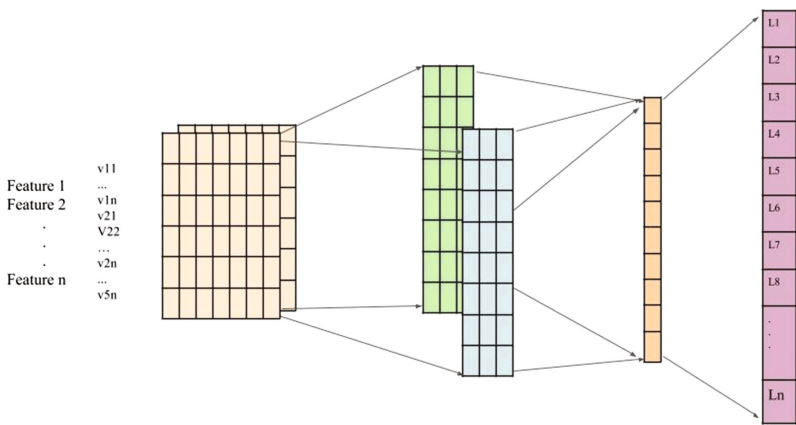


Fig. 1. Convolutional neural network architecture.

The conceptual representation of convolutional neural network is shown in Fig. 1, there are 4 main feature categories and hidden layers and an output layer which consists of $\mathbf{L}_1, \mathbf{L}_2, \dots, \mathbf{L}_n$ which are the locations detected, where n is equal to 2600.

Softmax regression is used to handle multiple classes. Here Y is the output and i is the number of samples.

$$Y(i) \in \{0, 1\}, \tag{1}$$

whereas in our scenario, softmax regression allows to handle 2600 classes.

$$Y(i) \in \{1, \dots, N\}, \tag{2}$$

where N refers to number of classes. The probability of the sum over the k possible output labels given input features $x(i)$ and the model parameter θ are calculated as:

$$P(y^{(i)} = k|x^{(i)};\theta) = \frac{\exp(\theta^{(k)T}x^{(i)})}{\sum_{j=1}^k \exp(\theta^{(j)T}x^{(i)})} \tag{3}$$

where P is the probability of getting a class y for the sample i , k is the number of possible classes. So, the cost function J is

$$J(\theta) = -\left[\sum_{i=1}^m \sum_{k=0}^1 \{y^{(i)} = k\} \log P(y^{(i)} = k|x^{(i)};\theta) \right] \tag{4}$$

Time Complexity of Convolutions is shown as follows, The total time complexity of all convolutional layers is: $O\left(\sum_{l=1}^d n_{l-1} \cdot s_l^2 \cdot n_l \cdot m_l^2\right)$.

Here l is the index of a convolutional layer, and d is the depth. n_l is the number of filters (also known as “width”) in the l -th layer. n_{l-1} is also known as the number of input channels of the l -th layer. s_l is the spatial size (length) of the filter. m_l is the spatial size of the output feature map.

4.4 Experimental Flow

Twitter’s streaming API is used for collecting data and store the data in the database. The collected data is iterated through for cleaning and pre-processing of each entry. For each location, location indicative terms are collected. After that extract the stylometric features and build model. See Fig. 2.

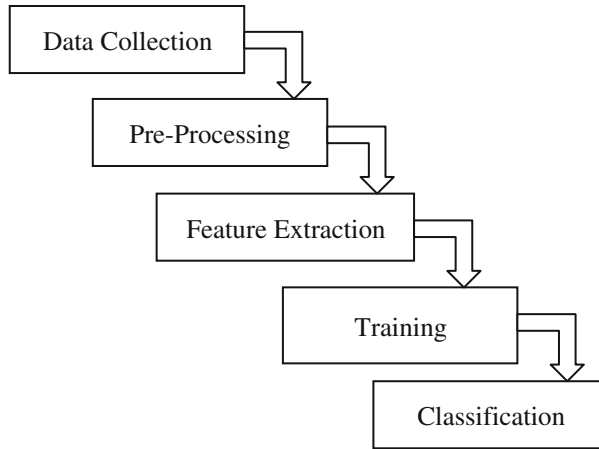


Fig. 2. Experimental flow.

5 Experimental Results

The stylometric features are extracted and different classification methods are used. This is the first time CNN has been employed to predict the geographic location on a dataset that is very huge with 100 Million tweets and outperforms other methods such as Random Forest - RF, Decision Tree - DT, Naive Bayes - NB with an accuracy of 51% with a distance error of 100 km. It's 54% and 59% respectively for the distance errors of 300 km and 600 km (Fig. 3). Along with this, the current method of finding location indicative terms are extended by considering the phrases also which helped in improving

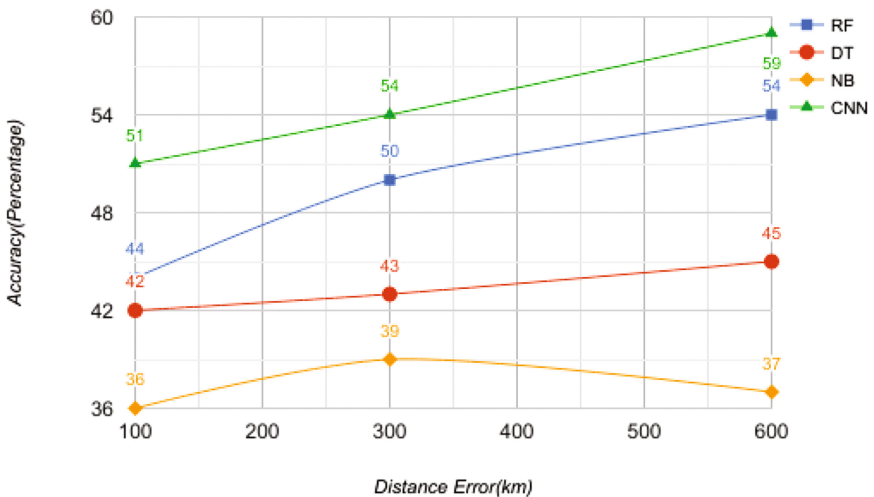


Fig. 3. Distance error vs. accuracy.

the prediction. By employing these two methods together accuracies of 55%, 58%, and 62% are achieved for distance errors of 100 km, 300 km and 600 km respectively. Approaches other than CNN yields less information gain comparatively as a result of over fitting and overlapping features because of higher dimension of classes [2600] and huge amount of dataset [100 million].

6 Conclusion

In this paper, we have presented a geographical location estimation method using stylistic based approach for social media users purely based on content of their posts. The results are further improved with location indicative terms. From our research, few observations were made where the writing style of people vary based on their geographical location. The geographic location prediction results are encouraging. Also this system can be deployed in such a way that it can adopt to the changes in the tweet writing pattern specific to region by actively rebuilding the model in real-time with the dataset collected. In the future work, more layers to the CNN model will be added to experiment and see for improvement in accuracy without affecting the scalability. Also this work will be extended to include Non-English languages that contains several dialects that increases the complexity significantly.

References

1. Twitter Usage Statistics: <http://www.internetlivestats.com/twitter-statistics/> (2017). Accessed 07 Apr 2017
2. Hutchinson, S.: Social media plays major role in Turkey protests. <http://www.bbc.com/news/world-europe-22772352> (2013). Accessed 08 Apr 2017
3. Kapko, M.: How social media is shaping the 2016 presidential election. <http://www.cio.com/article/3125120/social-networking/how-social-media-is-shaping-the-2016-presidential-election.html> (2016). Accessed 08 Apr 2017
4. Twitter: Adding your location to a Tweet. <https://support.twitter.com/articles/122236> (2017). Accessed 09 Apr 2017
5. Facebook: How do I add my location to a post? <https://www.facebook.com/help/115298751894487/> (2017). Accessed 10 Apr 2017
6. Facebook: How do I turn on Location Services for Facebook? https://www.facebook.com/help/275925085769221?helpref=faq_content (2017). Accessed 10 Apr 2017
7. Dialect Survey Maps and Results (2003). <http://dialect.redlog.net/maps.html>. Accessed 10 Apr 2017
8. Cheng, Z., Caverlee, J., Lee, K.: You are where you tweet: a content-based approach to geolocating twitter users. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp. 759–768. ACM (2010)
9. Chandra, S., Khan, L., Muhaya, F.B.: Estimating twitter user location using social interactions—a content based approach. In: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), pp. 838–843. IEEE (2011)
10. Bo, H., Cook, P., Baldwin, T.: Geolocation prediction in social media data by finding location indicative words. In: Proceedings of COLING, pp. 1045–1062 (2012)

11. Jurgens, D.: That's what friends are for: inferring location in online social media platforms based on social relationships. *ICWSM* **13**, 273–282 (2013)
12. McGee, J., Caverlee, J., Cheng, Z.: Location prediction in social media based on tie strength. In: *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, pp. 459–468. ACM (2013)
13. Kong, L., Liu, Z., Huang, Y.: Spot: locating social media users based on social network context. *Proc. VLDB Endow.* **7**(13), 1681–1684 (2014)
14. Compton, R., Jurgens, D., Allen, D.: Geotagging one hundred million twitter accounts with total variation minimization. In: *2014 IEEE International Conference on Big Data (Big Data)*, pp. 393–401. IEEE (2014)
15. Madden, M.: Privacy management on social media sites. In: *Pew Internet Report*, pp. 1–20 (2012)
16. Molok, N.N.A., Ahmad, A., Chang, S.: Information leakage through online social networking: opening the doorway for advanced persistence threats. *J. Aust. Inst. Prof. Intell. Off.* **19**(2), 38 (2011)
17. Shuyo, N.: Language detection library for Java. <https://github.com/shuyo/language-detection> (2010). Accessed 7 July 2016
18. Grierson, G.A. (ed.): *Linguistic Survey of India*, vol. 4. Office of the Superintendent of Government Printing, Calcutta (1906)
19. Stylometry: <https://en.wikipedia.org/wiki/Stylometry> (2017). Accessed 17 Apr 2017
20. Eisenstein, J., O'Connor, B., Smith, N.A., Xing, E.P.: A latent variable model for geographic lexical variation. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 1277–1287. Association for Computational Linguistics (2010)
21. Hecht, B., Hong, L., Suh, B., Chi, E.H.: Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 237–246. ACM (2011)
22. Rangel, F., Rosso, P., Koppel, M.M., Stamatatos, E., Inches, G.: Overview of the author profiling task at PAN 2013. In: *CLEF Conference on Multilingual and Multimodal Information Access Evaluation*, pp. 352–365. CELCT (2013)
23. Readability: <https://en.wikipedia.org/wiki/Readability> (2017). Accessed 20 Apr 2017
24. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint <http://arxiv.org/abs/1408.5882> (2014)
25. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors. In: *Proceedings of the 19th International Conference on World Wide Web*, pp. 851–860. ACM (2010)
26. Anil, R., Manjusha, K., Kumar, S.S., Soman, K.P.: Convolutional neural networks for the recognition of Malayalam characters. In: *Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014*, pp. 493–500. Springer (2015)
27. Athira, S., Mohan, R., Poornachandran, P., Soman, K.P.: Automatic modulation classification using convolutional neural network. *Int. J. Control Theory Appl.* **9**(16), 7733–7742 (2016)
28. Miura, Y., Taniguchi, M., Taniguchi, T., Ohkuma, T.: A simple scalable neural networks based model for geolocation prediction in Twitter. *WNUT 2016*, 9026924, 235 (2016)

EMG Pattern Classification Using Neural Networks

Tanmay Gupta^(✉), Jyoti Yadav, Shubham Chaudhary,
and Utkarsh Agarwal

Netaji Subhas Institute of Technology, Sector 3, Dwarka, Delhi 110078, India
tanmayg.ic@nsit.net.in, jyoti@nsit.ac.in,
shubhamchaudhary134@gmail.com, utkarsh1604@gmail.com

Abstract. The functioning of electromyogram (EMG) driven prosthesis to control the performance of artificial prosthetic arms placed on people with missing limbs depends on the cumulative effect of multiple dynamic factors, some of which include electrode placement position, muscle contraction levels, forearm orientations, etc. However, the study of the combined influence of these dynamic factors has been limited and hence offered us scope to improve the accuracy of the previous studies. We used the data to extract multiple features through the Time Dependent Power Spectrum Descriptor (TD-PSD) algorithm, which has proven to be one of the best methods of feature extraction. Samples are classified using the Neural Pattern Recognition Toolbox with scaled conjugate gradient backpropagation as the training algorithm, which gives an improved accuracy over Support Vector Machine (SVM) classifier. Neural Network is trained using the EMG signals of 10 subjects performing multiple hand movements to achieve classification accuracy up to 94.7%. The results obtained are a testimony to the fact that the suggested method is competent to improve the operation of pattern recognition myoelectric signals.

Keywords: Feature extraction · Pattern recognition · Clustering · Classification

1 Introduction

In a country like India, with more than half the population working in the agricultural or labor sector where physical injuries are inevitable, a large number of people undergo limb losses. Apart from the limb loss due to physical injuries, India is also the third largest home to diabetes patients who eventually have to amputate their limbs to the disease [1]. The discovery of an artificial limb or prosthesis provides the much needed respite to the amputees in India and across the globe.

Prosthesis is defined as a man-made device attached to the subject body to replace a part which could have been lost through trauma, disease or other congenital conditions. The statistics of the National Limb Loss Information Centre reveal that the upper limb amputations account for a much greater part of the overall trauma related amputations [2] and hence in this entire paper, we focus on the arm amputations and the prosthetic arms.

Many different kinds of robotic arms have been developed since its inception [3], ranging from grippers capable of performing basic industrial applications like lifting, moving, etc., to extremely detailed ones capable of enacting every possible human hand function.

The functioning of prosthetic arms is indeed a complex mechanism and involves deciphering the intended movement with the help of the electrical activities of the muscles, which are known as electromyogram (EMG) signals. The recording of the EMG signals is done through non-invasive methods through the surface of the skin of the amputee.

An overall process of controlling a prosthetic arm usually involves an electrode placed directly on the surface of the amputee used to record the signals, which are amplified, filtered and sampled to get a refined data set to be considered for deciphering the movement to be performed. This data is used for EMG pattern classification which includes processing of EMG signals, extraction of features and classification [4].

Continuous research in this field has brought advancements in this field with around 90% accuracy [5]. A number of factors that influence pattern recognition have been studied, for example, strength exerted by muscle [6], limb orientation [7], and electrode placement [8]. Other forms of noise may also cause problems [9]. The effect of these factors has been studied individually but [10] recently studied the effects of a combination of muscle contraction levels and forearm orientation on classification of EMG signals.

Feature extraction methods are used to get useful information out of almost meaningless and random time series EMG signals. Feature extraction can be done using methods like Time Domain based features [11], Discrete Fourier Transform [12], or Time Domain-Power Spectral Descriptors (TD-PSD) [13], etc. TD-PSD method as it has been established to be the best among all methods by previous literature [10]. TD-PSD method quantifies the angle of the EMG pattern rather than the amplitude which gives more robust results in comparison to other feature extraction methods. Feature extraction is then followed by a classifier which ultimately differentiates between actions and force levels with which it needs to be performed by giving an input to the digital controllers which controls the prosthetic arm. Support Vector Machine classifier was used in [10]. It has been proven to be equally good if not better to other classifiers such as Linear Discriminant Analysis (LDA), kNN, Random Forest and Naive Bayes [10].

This paper has tried to study the combined influence of different orientations of the forearm and varied muscular contraction levels on EMG pattern recognition. It has tried to continue and improve on the work done in this domain previously [10] and used their data set which was recorded live in Iraq as well as Australia. The data consists of ten intact limbed amputee subjects on which multiple electrodes were placed. These electrodes captured the EMG signals generated by 6 different movement classes at 3 varied contraction levels, with 3 trials given to each recording performed in 3 different orientation of the forearm. This data was used to extract the features by using the TD-PSD feature extraction method. Here, we have utilised the Neural Network Pattern Recognition toolbox available on MATLAB 2014 and subsequently compared the classification accuracy of this classification method with the previously used support vector machine classification method.

Section 2 discusses the data acquisition as well as the major methodologies and concepts used in feature extraction as well as classification involved in the study. Section 3 discusses the results obtained followed by the concluding remarks in Sect. 4.

2 Methods

2.1 Experimental Protocol

A normal computer display is placed in front of the subjects. The subjects were made to perform six actions or movements, which are, closed fist, opened fist, extension of wrist, flexion of wrist, wrist ulnar deviation, and wrist radial deviation. Three forearm orientations were considered: wrist fully supinated, at rest, and fully pronated, marked as 1st, 2nd, and 3rd orientations. Movements are performed at varied contraction levels: low, average and high in each orientation of all the six movements. Overall, each subject gave 162 trials: 6 classes of moves \times 3 orientations of the arm \times 3 levels of muscular contraction \times 3 trials per movement.

2.2 Feature Extraction Method: Time Domain Power Spectrum Descriptors

Recent studies in Electromyogram (EMG) pattern recognition show the error when the implementation of myoelectric control system is carried out [14]. When tests are carried out on EMG patterns for the same movement at different position, the controller shows limited performance. The feature extraction method is known as TD-PSD [13] is utilized to minimize the effect of limb position on classification.

The feature vector is obtained in two steps. In the first step, using the sampled time series EMG signal and transforming them through Fourier transform and Parseval's relations, a set of power spectrum features is extracted. Then the sampled time domain EMG signal is logarithmically scaled and the power spectrum moments are obtained from it. This is also called cepstral feature extraction method.

In the final step, the total six features are extracted which are nothing but the orientation between the power spectrum moments for original electromyogram signal and its cepstral version using a cosine similarity rule. The next section explains the feature extraction method in detail.

In Fig. 1 $x[j]$ with $j = 1, 2, \dots, N$, of length N denotes the sampled set of EMG signal. EMG trace within a certain epoch can be expressed as a function of frequency by means of Discrete Fourier transform (DFT). The feature extraction process begins by observing Parseval's theorem which states that the sum of the square of the function is equal to the sum of the square of its transform.

$$\sum_{j=0}^{N-1} |x[j]|^2 = \frac{1}{N} \sum_{k=0}^{N-1} |X[k]X^*[k]| = \sum_{k=0}^{N-1} |P[k]| \quad (1)$$

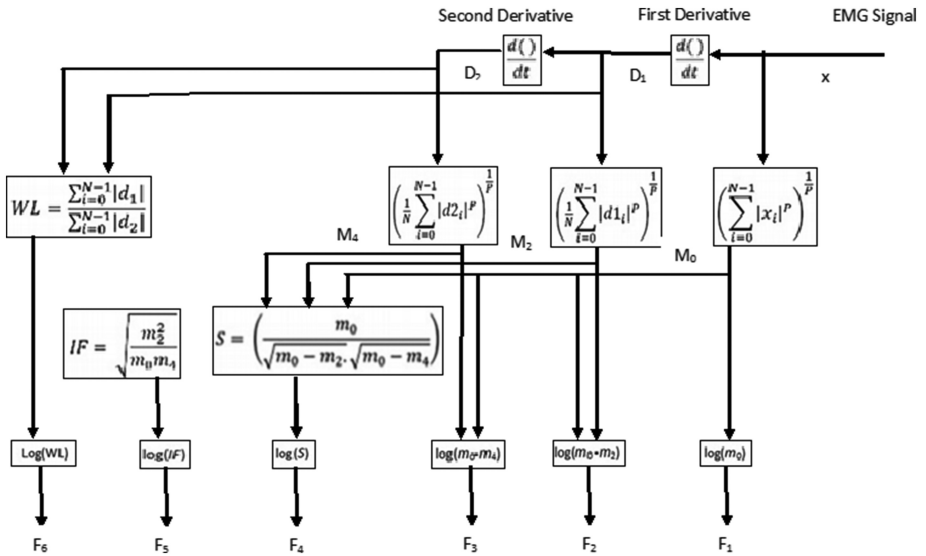


Fig. 1. Block diagram of time domain power spectrum descriptors feature extraction method

Frequency index is denoted by k and $P[k]$ is the power spectrum without phase [15]. This method will deal with the whole spectrum because the full frequency description is symmetric in nature which we obtained from Fourier Transform and we cannot obtain the power spectral density directly from the time-domain. So, all odd moments will be considered as zero. So, m is denoted as moment and n as order of the moment of the power spectrum $P[k]$.

$$m_n = \sum_{k=0}^{N-1} k^n P[k] \tag{2}$$

In the equation shown above, if the value of n is nonzero then the Fourier transform’s time-differentiation property is used and when $n = 0$ then Parseval’s theorem is used. This kind of property states that Δ^n is denoted as discrete time signals, which can also be written as multiplying the $X[K]$ by k to the n^{th} power.

$$F[\Delta^n x[j]] = k^n X[k] \tag{3}$$

The moments which will help in extraction of feature sets as shown in Fig. 1 are:

Root squared zero order moment: This feature mainly denotes the strength of muscle contraction, or the frequency-domain’s total power.

$$\bar{m}_0 = \sqrt{\sum_{j=0}^{N-1} x[j]^2} \tag{4}$$

Root squared second order moments: Power spectrum is denoted as the second order moments.

$$\bar{m}_2 = \sqrt{\frac{1}{N} \sum_{j=0}^{N-1} (\Delta x[j])^2} \quad (5)$$

Root squared fourth order moments: For the fourth order moment we raise the power of frequency index by 2 in second order moment.

$$\bar{m}_4 = \sqrt{\frac{1}{N} \sum_{j=0}^{N-1} (\Delta^2 x[j])^2} \quad (6)$$

So, the second and the fourth moment derivatives of the signals are used to minimize the signal's full energy; hence, we normalize ($\lambda = 0.1$) to limit the influence of noise on all moments.

$$m_0 = \frac{\bar{m}_0}{\lambda} \quad m_2 = \frac{\bar{m}_2}{\lambda} \quad m_4 = \frac{\bar{m}_4}{\lambda} \quad (7)$$

Now, the first three features using the above moments are:

$$f_1 = \log(m_0) \quad (8)$$

$$f_2 = \log(m_0 - m_2) \quad (9)$$

$$f_3 = \log(m_0 - m_4) \quad (10)$$

The other three features are:

Sparseness: Sparseness measures the amount of energy is packed in only minor components of a vector. This feature can be expressed as:

$$f_4 = \log\left(m_0 / \sqrt{(m_0 - m_2)(m_0 - m_4)}\right) \quad (11)$$

Such a feature describes a vector with all elements equal with a sparseness measure of zero that is m_2 and m_4 equal to zero because of differentiation and so $f_4 = 0$. For all other sparseness levels, value should be greater than 0 [14].

Irregularity Factor (IF): It denotes the measure of ratio of the count of upward zero crossings to the number of peaks. This feature can be expressed as in terms of spectral moments:

$$f_5 = \log(m_0 / \sqrt{m_0 m_4}) \quad (12)$$

Waveform Length Ratio (WL): The summation of absolute value of the first and second derivative of EMG signal over its entire length is calculated to find the waveform length (WL) feature using the formula:

$$f_6 = \log \left(\frac{\sum_{j=0}^{N-1} |\Delta x|}{\sum_{j=0}^{N-1} |\Delta^2 x|} \right) \quad (13)$$

Now on the basis of Fig. 1, we form a matrix $a = [a_1, a_2, a_3, a_4, a_5, a_6]$ by using the six extracted features. We also add another feature vector, expressed as $b = [b_1, b_2, b_3, b_4, b_5, b_6]$, which is extracted logarithmically scaled version $\log(x^{2x^2})$. For each EMG channel the final 6 features are extracted which are nothing but the orientation between the vectors obtained earlier. A cosine similarity rule is used to find these features given as

$$f_i = \frac{-2a_i b_i}{a_i^2 + b_i^2} \quad (14)$$

2.3 Artificial Neural Network

It makes a network of artificial neurons that map the input to the output, both of which are known to us with certainty. A backpropagation network consists of at least three layer which are one input layer, one output layer, and one or more hidden layers in between the input and output layers. The hidden layer has a number of hidden neurons. The network is trained, that is, the various connection weights and bias values are adjusted so as to generate the desired outputs for the given inputs. The error generated at the output is the difference in our desired output and the present output at the output nodes. This error is backpropagated from the output layer to the input layer through the hidden layer(s). This changes the connection weights to reduce the error. This process is called backpropagation.

For the classification process, Neural Pattern Recognition toolbox has been used, available in MATLAB 2014, where the extracted TD-PSD features from EMG signals obtained from 10 subjects are treated as input, that is, the data to be classified. Networks of pattern recognition are feed-forward networks which can be used to train and classify inputs according to their target classes.

To train the network, we use the scaled conjugate gradient backpropagation method of classification. This training function comprises of three layers in total. There is just one hidden layer. The number of neurons is selected according to our network size.

3 Experimental Results

To report our results, we combine all the data and classification results from our 10 subjects. There are a total of 1620 samples from 162 trials of our 10 subjects. We first divided our samples randomly to feed into the pattern recognition tool using scaled conjugate gradient backpropagation algorithm. We used 70% of those samples for training, 15% of those samples for validation and 15% of those samples for testing. Figure 2 shows the confusion matrix for training with TD-PSD features:

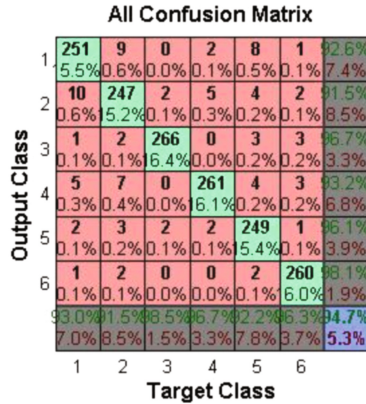


Fig. 2. Preliminary testing results with the entire dataset.

We now conducted a systematic study where, at each forearm orientation, we selected the data from one contraction level to train our classifier and then used the data from all contraction levels for testing our model.

For the case when we use the data of low contraction level in orientation 1 for training, Fig. 3 shows the confusion matrix and receiver operating characteristics (ROC), and Fig. 4 shows the performance curve. These plots help in the gauging the efficiency of a supervised learning algorithms (Table 1).

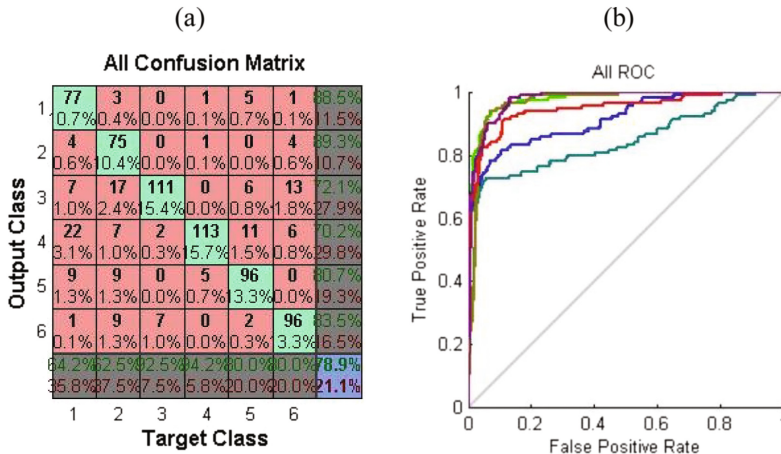


Fig. 3. (a) Confusion matrix and (b) ROC for training with low force level at orientation 1 and testing with all forces of the same orientation.

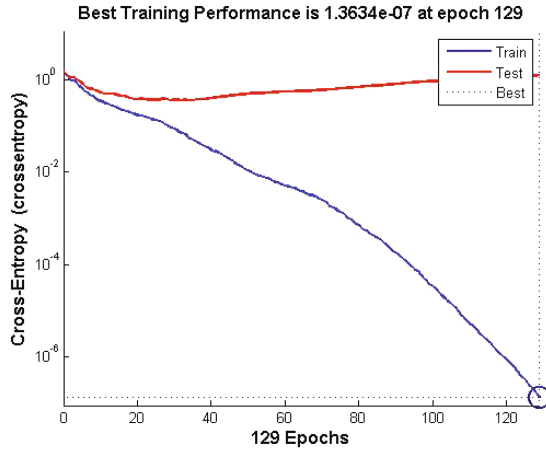


Fig. 4. Performance curve for training with low force level at orientation 1 and testing with all forces of the same orientation.

Table 1. Same orientation analysis. Testing was done using all forces within an orientation.

Orientation	Training and testing parameters	Accuracy %
Orientation 1	Train Low Force Hidden Neurons = 140	78.9%
	Train Medium Force Hidden Neurons = 140	85.4%
	Train High Force Hidden Neurons = 140	78.2%
Orientation 2	Train Low Force Hidden Neurons = 47	81.8%
	Train Medium Force Hidden Neurons = 35	83.6%
	Train High Force Hidden Neurons = 35	79.3%
Orientation 3	Train Low Force Hidden Neurons = 47	79.4%
	Train Medium Force Hidden Neurons = 35	81.4%
	Train High Force Hidden Neurons = 35	80.4%

The classification accuracy was the highest when training data from medium contraction level was used, at each forearm orientation.

Next, the classifier was trained and tested with data from different orientations. Specifically, for training, each orientation was selected one by one and the TD-PSD features extracted from data of all three muscle contraction levels was used. The TD-PSD features exacted from all the contraction levels of the other two orientations was the testing data.

There was a decline in classification performance from the earlier scenario, predictably so. For the case when the network was trained using data from orientation 1, and tested using data from orientations 2 and 3, Fig. 5 shows the confusion matrix and ROC. Figure 6 shows the performance curve.

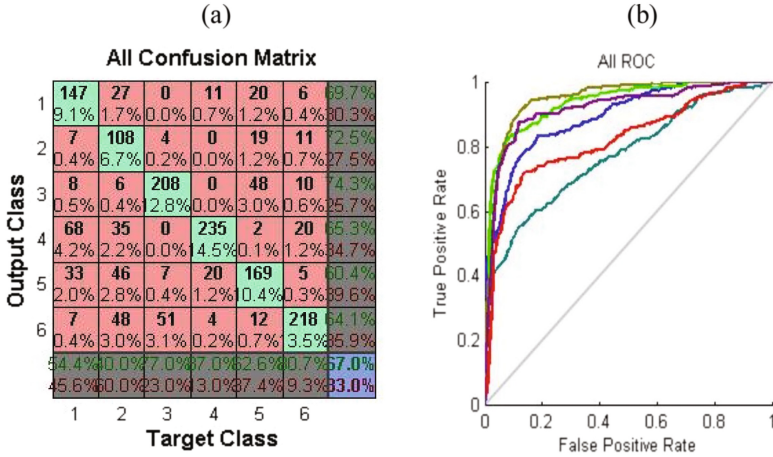


Fig. 5. (a) Confusion matrix and (b) ROC for training the classifier using orientation 1 data. Testing using data from orientation 2 and 3

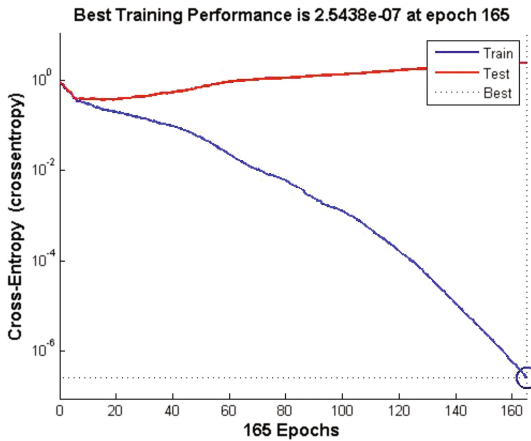


Fig. 6. Performance curve for training the classifier using orientation 1 data. Testing using data from orientation 2 and 3

Training the classifier with data from orientation 2 (as opposed to orientations 1 or 3) and testing the network with the other orientations resulted in the highest classification accuracy, which is 69.1% on average.

These results can be compared to those obtained by using support vector machine (SVM) classifier with SVM parameters as $C = 32$ and $\gamma = 0.0625$ (Table 2).

Table 2. Different orientation analysis.

Training and testing data	Movement class	Neural network classification accuracy	Support vector machine accuracy
Train: Orientation 1 Test: Orientation 2 and 3	C1	69.7%	66%
	C2	72.5%	50%
	C3	74.3%	60%
	C4	65.3%	55%
	C5	60.4%	48%
	C6	64.1%	63%
Train: Orientation 2 Test: Orientation 1 and 3	C1	66.9%	70%
	C2	67.0%	37%
	C3	86.5%	68%
	C4	77.3%	57%
	C5	63.7%	62%
	C6	67.7%	62%
Train: Orientation 3 Test: Orientation 1 and 2	C1	68.8%	65%
	C2	50.8%	38%
	C3	71.2%	75%
	C4	77.2%	35%
	C5	60.0%	61%
	C6	63.9%	57%

4 Concluding Remarks

Using TD-PSD as the feature extraction method, a comprehensive study of how varying muscle contraction levels and different forearm orientations together affect the EMG pattern recognition was conducted. It has been well established that the TD-PSD is a superior feature extraction method in comparison to other feature extraction methods. This paper has employed a new classifier, namely, Neural Network Classifier with scaled conjugate gradient back propagation training algorithm on MATLAB to improve the classification accuracy. Training this new classifier with data of six movement classes, each carried out with multiple muscular contraction levels and at various forearm orientations, by using TD-PSD features offers satisfactory classification accuracy and an improvement over Support Vector Machine classifier.

Maximum classification accuracy is obtained when the training data includes all forearm orientations. It is also observed that using training data from medium muscle contraction level provides the best accuracy when testing in comparison to low or high force levels.

References

1. Nazarpour, K., Sharafat, A.R., Firoozabadi, S.M.P.: Application of higher order statistics to surface electromyogram signal classification. *IEEE Trans. Biomed. Eng.* **54**(10), 1762–1769 (2007)
2. Owings, M.F., Kozak, L.J.: Ambulatory and inpatient procedures in the United States, 1996. *Vital Health Stat.* **139**, 1–119 (1998)
3. Belter, J.T.: Mechanical design and performance specifications of anthropomorphic prosthetic hands: a review. *J. Rehabil. Res. Dev.* **50**(5), 599 (2013)
4. Boostani, R., Moradi, M.H.: Evaluation of the forearm EMG signal features for the control of a prosthetic hand. *Physiol. Meas.* **24**(2), 309 (2003)
5. Rasool, G., et al.: Real-time task discrimination for myoelectric control employing task-specific muscle synergies. *IEEE Trans. Neural Syst. Rehabil. Eng.* **24**(1), 98–108 (2016)
6. Al-Timemy, A.H., et al.: A preliminary investigation of the effect of force variation for myoelectric control of hand prosthesis. In: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE (2013)
7. Peng, L., et al.: Combined use of semg and accelerometer in hand motion classification considering forearm rotation. In: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE (2013)
8. Hargrove, L., Englehart, K., Hudgins, B.: A training strategy to reduce classification degradation due to electrode displacements in pattern recognition based myoelectric control. *Biomed. Signal Process. Control* **3**(2), 175–180 (2008)
9. Spanias, J.A., Perreault, E.J., Hargrove, L.J.: Detection of and compensation for EMG disturbances for powered lower limb prosthesis control. *IEEE Trans. Neural Syst. Rehabil. Eng.* **24**(2), 226–234 (2016)
10. Khushaba, R.N., et al.: Combined influence of forearm orientation and muscular contraction on EMG pattern recognition. *Expert Syst. Appl.* **61**, 154–161 (2016)
11. Hakonen, M., Piitulainen, H., Visala, A.: Current state of digital signal processing in myoelectric interfaces and related applications. *Biomed. Signal Process. Control* **18**, 334–359 (2015)
12. He, J., et al.: Invariant surface EMG feature against varying contraction level for myoelectric control based on muscle coordination. *IEEE J. Biomed. Health Inf.* **19**(3), 874–882 (2015)
13. Al-Timemy, A.H., et al.: Improving the performance against force variation of EMG controlled multifunctional upper-limb prostheses for transradial amputees. *IEEE Trans. Neural Syst. Rehabil. Eng.* **24**(6), 650–661 (2016)
14. Khushaba, R.N., et al.: Towards limb position invariant myoelectric pattern recognition using time-dependent spectral features. *Neural Netw.* **55**, 42–58 (2014)
15. Khushaba, N., et al.: A fusion of time-domain descriptors for improved myoelectric hand control. In: 2016 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE (2016)
16. Bhardwaj, N., et al.: Extraction of EMG signal in a software compatible format from an online database using WFDB package. *Persp Sci.* **8**, 767–769 (2016)
17. Agarwal, S., et al.: EEG signal enhancement using cascaded S-Golay filter. *Biomed. Signal Process. Cont.* **36**, 194–204 (2017)

Crime Against Women: A State Level Analysis Using a Hierarchical and K-Means Clustering Techniques

Ayushi Dhawan and M.G. Deepika^(✉)

Department of Management, Amrita University, Bangalore, India
mgdeepika@gmail.com

Abstract. Despite the constitutional and legal provisions, crime against women in India has been continuously on the rise. While the incidence of crime against women is sharply on the rise, different dimensions of crime have varied across the states. Given the NCRB data on crime against women in India at the state level, the current study examines the dimensions and changing trends on crime against women and its association with important socio economic variables. Cluster analysis using agglomerative hierarchical method suggests for three clusters. We then use the widely used K-Means clustering method to arrive at the final clusters. The analysis reveals strong association of the rate of crime with that of socio economic variables like poverty rate, per capita state domestic product, literacy rate and the human development ranking of the state. The study reveals the importance of development indicators as much as the legal provisions in bringing down the rate of crime against women in India.

Keywords: Crime · Women · India · Development · Cluster analysis

1 Introduction

Despite constitutional provisions and legal advancements to prevent and deal with crime against women, the rate of this crime in India is continuously on the rise. The incidence of crime against women has more than doubled over the past ten years [1]. NCRB report reveals that crime has been recorded against women in every three minutes which could be in the form of rape, dowry death, suicide or domestic violence [1]. Women in India continue to suffer from violence of different depths and dimensions.

While India has progressed in economic growth since economic liberalization in the nineties, it struggles to guarantee the levels of freedom for women, where women in India are not yet totally liberated from the dominance of men. The Thompson Reuters Survey of G 20 nations in 2012 shows India ranks in the worst place to be a woman [2]. In addition, gender based violence including rape, domestic violence, mutilation and sexual abuse are seen as serious causes of health problem for women [3]. At a global

level the health burden from gender based victimization is comparable to that from other conditions already high on the world agenda. Addressing this issue to understand the nature and causes for crime against women and the laws to prevent and act effectively is a matter of utmost importance.

While the incidence of crime against women in India is sharply on rise, different dimensions of crime have varied across the states [1]. NCRB report in 2015 highlights the alarming increase in crime in some of the major cities of India, the NCR region of Delhi being on the top [4]. The causes for crime against women can be broadly classified under economic, social, demographic, psychological and legal [5]. While factors like unemployment, poverty, education, male economic dominance are the economic causes; power imbalance, female dependency, alcohol consumption, sex ratio are the social and demographic factors. Personality disorders, low self esteem, lack of assertiveness, lack of moral education are the psychological factors. Inefficiency of laws can be a legal cause leading to high crime rate.

Studies earlier have examined the influence of some of the socio economic factors on crime against women of specific nature like rape and domestic violence with specific reference to India [6, 7]. Analysis of overall crime at a global level across and cities and states of countries have been conducted earlier using the multivariate analysis largely with classification techniques [8–10]. Given the NCRB data on crime against women in India at the state level, the current study examines the dimensions and changing trends on crime against women and its association with important Socio economic variables at a state level.

2 Related Work

We come across very few studies specifically analyzing the causes of crime against women in India. A study by Dutta et al. examines the risk factors for domestic violence at the state level using the NFHS-3 database [6]. The paper attempts to examine many risk factors specifically with domestic violence. A limited dependent model is used to estimate the effects. A broad investigation revealed that men's exposure to intergenerational violence and alcohol consumption by partner emerge as significant risk factors. Other variables included were educational attainment by female as well as the husband, her employment status, the employment status of the husband, and demographic factors like caste and religion. Study by Mitra and Singh, 2007 shows the typical paradox in the state of Kerala when it comes to domestic violence committed against women [7]. The study shows that high educational attainment has fostered new aspirations and attitudes among women in Kerala. The asymmetry between the attitude of educated wife who is not willing to be typically a submissive wife and the conservative attitude of less educated husband who is perceived by societal norms to be the decision maker of the house often contributes to domestic violence. This kind of

domestic conflict between the husband and wife is generally absent in states where women are for the most part much less educated and more submissive to men.

Studies pertaining to analysis of overall crime at a global level, testing the association of crime rate with socio economic variables have largely used the classification techniques. Way back in 1970s Downs had demonstrated using factor analysis that cities in the US experiencing these violent events tend to be the largest, least rapidly growing, most densely populated whose citizens are less educated, have low income and higher unemployment rates than those in cities in which no incidents of racial violence have taken place [11].

Land et al. (1990) apply the discriminant factor analysis to group US cities with a high or low crime rate [8]. The variables that are included in the study are median family income, the percent of family living below poverty line, Gini index of family income inequality, percent of population that is black, and percent of children under age 18 not living with both parents. Williams and Gedeon (2004) have used multi-variate analysis to help classify US cities as safe or unsafe according to several variables [9]. Chiricos (1987) addressed the issue of a relationship between unemployment and crime rate and concluded that it is a positive one [10]. Research by economists in the 1960s has demonstrated that difference in socio-economic structure may result in or be the motivation for criminal acts. The work of Becker, Fliesher and Singell have demonstrated that variables which are surrogate measures of opportunity and price structure (employment, education, income and so on) may explain allocations of time between legitimate and illegitimate criminal activity [12–14].

3 Data Source and Definition of Variables

The study is purely based on secondary sources of information. Data on total incidence of crime and rate of crime against women in different states of India has been gathered from the National Crime Records Bureau for the years from 2000 to 2013 [1]. Data for other variables used for analysis like Sex ratio [15, 16], Literacy rate [17, 18], Poverty rate [19], Per capita SDP [20], HDR ranking of states [21], and unemployment rate [22] have been gathered from different government sources for the two census years of 2001 and 2011 and the unemployment rate available for the year 2015.

Total incidence of crime (TC) is the total number of incidences of crime against women reported each year. Rate of crime (RC) is the crime against women reported per lakh of population. Sex ratio (SR) is defined as the number of females per thousands of males. Literacy rate (LR) the total percentage of the population at a particular time aged seven years or above who can read and write.

Per capita SDP, (PSDP) is the Gross State domestic product divided by population of the state. Poverty Rate (PR) is defined as the ratio of the number of people (in a given age group) whose income falls below the poverty line; taken as half the median household income of the total population). Human Development Index and Rank

(HDI) (It is a composite statistic of life expectancy, education, and per capita income indicators, which are used to rank states Unemployment Rate (UR), is share of the labor force that is jobless, expressed as a percentage.

4 Methodology

The study has used descriptive statistics to understand the nature, dimensions and magnitude of different types of crimes against women across the states of India. To examine the association of the rate of crime with the select socio economic indicators we together use the hierarchical and K-means clustering techniques.

Cluster analysis is a technique used to classify objects or cases into relatively homogeneous groups called clusters. The objects in each cluster tend to be similar to each other and dissimilar to other clusters. Cluster Analysis is also called Numerical Taxonomy. Hierarchical cluster is a procedure characterized by the development of hierarchy or tree like structure. In Hierarchical Clustering procedure each object starts out in separate cluster. In this the clusters are formed by grouping objects into bigger and bigger clusters. By using the Agglomeration schedule the objects or cases are being combined at each stage of hierarchical clustering process. The Agglomeration schedule with the Dendrogram helps in identifying the right number of cluster to classify objects. Euclidian distance which is the square root of the sum of squared differences in values for each variable is the distance measure used. Linkage method using the Ward's procedure is used to arrive at linking the objects in the cluster. Ward's Procedure is a variance method in which the squared Euclidean distance to the cluster means is minimized. Once the right number of clusters are identified from the agglomeration schedule using the hierarchical clustering process, we then use the non-hierarchical K-Means clusters to identify the cluster membership. The non-hierarchical clustering method is frequently referred to as K-means clustering. In the sequential threshold method, a cluster center is selected and all objects within a pre-specified threshold value from the center are grouped together. Then a new cluster center or seed is selected and the process is repeated for the unclustered points. Once an object is clustered within the seed it is no longer considered for clustering with subsequent seeds.

K-Means is the most popular of the clustering algorithms. However, using of K-means needs the prior knowledge of the number of clusters. In the absence of the theoretical knowledge of the number of clusters we use the hierarchical agglomerative cluster which helps in arriving at the most suitable number of clusters denoted by the agglomeration schedule and the dendrograms. Once the ideal number of clusters is specified we then use the K-means to arrive at the final clusters. The K-means algorithm assigns each point to the cluster whose center is the nearest.

5 Results and Discussion

5.1 Patterns and Changing Trends on Crime Against Women

Crime against women in India are broadly classified by NCRB under two heads; Crimes under the Indian Penal Code (IPC) and Crimes under Special and Local Laws (SLL). Crimes under IPC includes Rape (Sec 376 of IPC), Kidnapping and Abduction (Sec 363–373 of IPC), Dowry Deaths (Sec 304-B of IPC), Physical and Mental Torture (Sec 498 of IPC), Molestation (Sec 354 of IPC), Sexual Harassment (Sec 509 of IPC) and Importation of Girls (Sec 366-B IPC). Government usually after regular interval of time review and make alteration if needed in the already existing laws that affect women. There are specific acts which have special provisions to safeguard women and their interests. Those are the crimes that fall under the SLL. These include - The employee state insurance Act of 1948; The Plantation Labor Act; The family Court Act, 1954; The special marriage Act, 1954; The Hindu marriage Act, 1955; The Succession Act, 1956; Immoral Traffic (Prevention) Act, 1956; The maternity Benefit Act, 1961; Dowry Prohibition Act, 1961; The medical termination of pregnancy Act, 1971; The contract labor Act, 1976; The equal Remuneration Act, 1976; The child marriage Restraint Act, 1979; The criminal law Act, 1983; The factories Act, 1986; Indecent Representation of women Act, 1986; Commission of Sati (Prevention) Act, 1987; and Domestic Violence Act, 2005.

Table 1. Incidence and rate of crime against women in India.

Year	I	R
2000	141373	14.1
2001	143795	14
2002	147678	14.1
2003	140601	13.2
2004	154333	14.2
2005	155553	14.1
2006	164765	14.7
2007	185312	16.3
2008	195856	17
2009	203804	17.4
2010	213585	18
2011	228650	18.9
2012	244270	41.74
2013	309546	52.24
2014	337922	56.3
2015	327394	53.9

Note: I = Incidence of crime, R = Rate of crime

Source: Table compiled by the author using data from NCRB from the year 2000 to 2015, <http://ncrb.nic.in/>

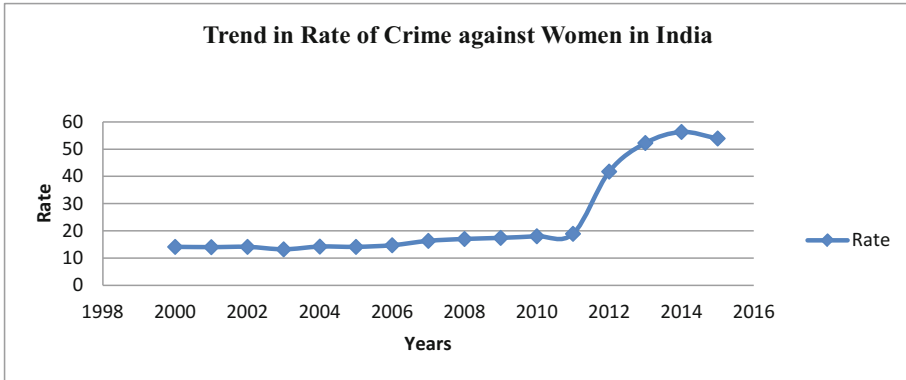


Fig. 1. Trends in rate of crime against women in India. **Note:** Rate of crime: The rates are calculated by National Crime Records Bureau as the number of incidents per 1,00,000 of population. **Source:** Figure compiled though the data gathered from NCRB data, various years, <http://ncrb.nic.in/>.

Table 1 along with Fig. 1 clearly shows the increase in the overall rate of crime in the recent years with a sharp increase since the year 2012. The overall crime rate against women has increased from 14.1 percent in the year 2000 to 53.9% in 2015. Figure 1 shows that there has been increase in crime rate from 2011 to 2015. The major states and cities that contribute to increase in crime rate are Delhi, Kerala, Haryana and Rajasthan. Kerala inspite of its high development indicators has reported overall 455 crimes on per lakh population. State like Haryana which has low sex ratio has also showed increase in crimes like female foeticide, early marriages, cruelty by husbands (Table 2).

The highest contributor to the total crime is that of Cruelty by husband and relatives (38.51) followed by Molestation (21.83), Kidnapping and Abduction (12.65) and Rape (11.05). Looking into the individual trends in the total crime over the years there is an overall increase on the rate of crime caused by cruelty by husband and relatives, kidnapping and abduction and rates in other crimes more or less remaining the same and eve teasing dropping consistently.

From Table 3 it is evident that among the states and cities of India, Delhi, Tripura, Assam, Andhra Pradesh, West Bengal, Rajasthan, Bangalore, Madhya Pradesh, Kerala, Haryana are with very high average rate of crime. Whereas Nagaland, Meghalaya, Manipur, Kolkata, Sikkim, Mumbai show low average rates of crime. A Sharp increase in percentage of crime rate is recorded in West Bengal, Meghalaya, Tripura, Assam, Sikkim, Kolkata, Kerala, Manipur and Delhi. States and cities like Chennai, Tamil Nadu, Arunachal Pradesh, Chhattisgarh, Uttar Pradesh, Madhya Pradesh, Punjab and Himachal Pradesh have shown a decline in crime rate.

Table 2. Percentage of different crimes to overall crime on women

Year	Rape	Kidnapping and abduction	Dowry deaths	Cruelty by husband and relatives	Molestation	Eve teasing	Importation of girls	Sati prevention act	Immoral traffic	Indecent representation of women	Dowry prohibition act
2000	11.67	9.87	4.86	32.27	23.30	7.80	0.05	0.00	6.73	0.47	2.03
2001	11.18	10.18	4.76	34.19	23.73	6.78	0.08	0.00	6.12	0.73	2.24
2002	11.09	9.82	4.62	33.34	22.98	6.88	0.05	0.00	7.61	1.70	1.91
2003	11.27	9.46	4.42	36.06	23.43	8.77	0.03	0.00	3.92	0.74	1.91
2004	11.81	10.09	4.55	37.66	22.40	6.48	0.06	0.00	3.72	0.89	2.33
2005	11.80	10.13	4.36	37.49	21.97	6.42	0.10	0.00	3.80	1.88	2.06
2006	11.74	10.57	4.62	38.31	22.22	6.05	0.04	0.00	2.76	0.95	2.73
2007	11.19	11.02	4.37	40.97	20.90	5.91	0.03	0.00	1.93	0.65	3.03
2008	10.96	11.71	4.17	41.53	20.63	6.24	0.03	0.00	1.36	0.52	2.84
2009	10.50	12.63	4.11	43.94	18.99	5.40	0.02	0.00	1.21	0.41	2.77
2010	10.38	13.95	3.93	44.03	19.01	4.66	0.02	0.00	1.17	0.42	2.43
2011	10.59	15.55	3.77	43.36	18.79	3.75	0.03	0.00	1.06	0.20	2.89
2012	10.20	15.66	3.37	43.61	18.57	3.76	0.02	0.00	1.05	0.06	3.70
2013	10.89	16.76	2.61	38.40	22.85	4.07	0.01	0.00	0.83	0.12	3.46
2014	10.87	16.96	2.50	36.36	24.34	2.88	0.00	0.00	0.61	0.01	2.97
2015	10.58	18.11	2.33	34.64	25.18	2.65	0.00	0	0.74	0.01	3.02
Average	11.05	12.65	3.96	38.51	21.83	5.53	0.04	0.00	2.79	0.61	2.65

Note: Percentage of each crime is calculated by dividing the incident of each crime to the total number of incidence of a particular year.

Source: The above data is compiled into a table by the author, by collecting the data from NCRB, <http://ncrb.nic.in/>. Each column consists of percentage of different crimes in each year.

Table 3. Rate of crime against women across the states and major cities of India (average rate of crime for the years 2000 to 2013), by NCRB

States	RC	State	RC
Delhi	37.58	Maharashtra	16.76
Tripura	35.84	Gujarat	16.48
Assam	35.15	Karnataka	15.61
Andhra Pradesh	33.13	Tamil Nadu	14.29
West Bengal	31.78	Jharkhand	13.51
Rajasthan	30.08	Uttarkhand	13.24
Bangalore	28.49	UP	12.82
Madhya Pradesh	27.29	Punjab	12.48
Kerala	25.88	Goa	11.47
Haryana	25.85	Bihar	10.44
Jammu & Kashmir	25.40	Mumbai	10.23
Orissa	23.64	Sikkim	10.16
Chhattisgarh	22.75	Kolkata	9.41
Chennai	18.92	Manipur	8.83
Arunachal Pradesh	18.05	Meghalaya	8.37
Mizoram	17.74	Nagaland	2.19
Himachal Pradesh	16.83	All India Average	24.39

Note: The above table consists of average of Total Rate of Crime (2000–2013) of each state and city. They are arranged in descending order on the basis of All India rate average.

Source: The above table is compiled by the author, using the relevant data about the states from, <http://ncrb.nic.in/>.

5.2 Results of Cluster Analysis

To examine the association of rate of crime with socio economic variables we conduct cluster analysis using the hierarchical and K-means clustering techniques. The variables included in the analysis are the rate of crime, the total incidence of crime, sex ratio, literacy rate, poverty rate, percapita State Domestic Product, Human development rank and unemployment rate in the state level. The Agglomeration schedule and the dendrogram (shown in Fig. 2) indicate three clusters as ideal. We then use the K-Means cluster analysis indicating three clusters to arrive at cluster membership and the final cluster centers showing the mean values of the variables (Tables 4, 5 and 6).

The results show 13 states in cluster 1, 14 in cluster 3 and one in cluster 2. The mean values of rate of crime and the total incidence of crime in states cluster 1 and 2 is lower compared to cluster three. The states in cluster 1 and 2 on an average show significant association with high literacy rate, low poverty rate, high per capita SDP and good Human development rankings. However, there is no association found with unemployment and sex ratio. The state of Goa turns out to be a classic case of good development indicators with respect to most of the variables in the analysis and low

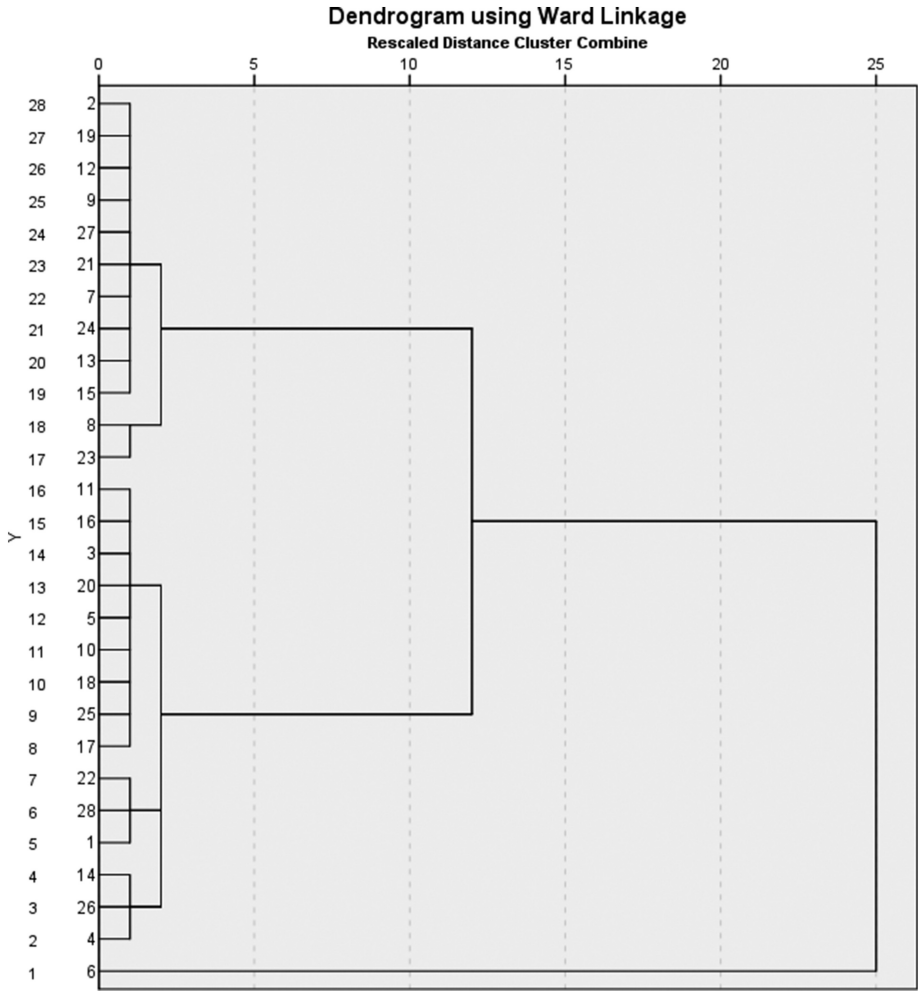


Fig. 2. Dendrogram showing the ideal number of clusters

Table 4. Number of cases in each cluster

Cluster	No. of cases
1	13
2	1
3	14

Table 5. Cluster membership (states under each cluster)

Cluster 1	Cluster 2	Cluster 3
Andhra Pradesh	Goa	Assam
Arunachal Pradesh		Bihar
Gujarat		Chhattisgarh
Haryana		Jammu & Kashmir
Himachal Pradesh		Jharkhand
Karnataka		Madhya Pradesh
Kerala		Manipur
Maharashtra		Meghalaya
Nagaland		Mizoram
Punjab		Orissa
Sikkim		Rajasthan
Tamil Nadu		Tripura
Uttarkhand		UP
		West Bengal

Note: This table shows the position of each state that has been created by the author on the basis of result obtained by doing K means cluster analysis.

Table 6. Final cluster centers

	Cluster		
	1	2	3
RA	17.00	11.47	21.69
TC	6053.86	142.35	7417.81
SR	942	965	944
LR	75	85	69
PR	16	5	28
PCSDP	57220	144269	30777
HDR	11	2	19
UR	68	69	63

Note: RC (Rate of crime), TC (Total incidence of crime) SR - Sex Ratio; LR - Literacy Rate; PR - Poverty Rate; PCSDP - Per capita State Domestic Product; HDR - Human Development Rank; UR - Unemployment Rate.

crime rate. Some states like Kerala, Manipur and Meghalaya are misclassified when looked at classification based on total rate of crime and incidence of crime. This paradox can however be explained. Kerala inspite of high development indicators has been experiencing high crime against women due to asymmetry between the education status of male and female. Kerala also records high rate of crime largely due to high reporting by the victimized women. Tribal states of north east like Manipur and

Meghalaya has very low rate of crime against women in spite of low development indicators largely due to matrilineal family structure where a woman is the head of the family.

6 Conclusion

While the incidence of crime against women in India is sharply on rise, different dimensions of crime have varied across the states. To examine the association of rate of crime with socio economic variables we conduct cluster analysis using the hierarchical and K-means clustering techniques. The variables included in the analysis are the rate of crime, the total incidence of crime, sex ratio, literacy rate, poverty rate, percapita SDP, Human development rank and unemployment rate in the state level. The Agglomeration schedule and the dendrogram indicate three clusters as ideal. We then use the K-Means cluster analysis to arrive at cluster membership and the final cluster centers showing the mean values of the variables. The mean values of rate of crime and the total incidence of crime in states cluster 1 and 2 is lower compared to cluster three. The states in cluster one and two on average show significant association with high literacy rate, low poverty rate, high percapita SDP and good Human development rankings. However, there is no association found with unemployment and sex ratio. The state of Goa turns out to be a classic case of good development indicators with respect to most of the variables in the analysis and low crime rate. Some states like Kerala, Manipur and Meghalaya are misclassified when looked at classification based on total rate of crime and incidence of crime. The study reveals the importance of development indicators like literacy, economic growth, lifting population above poverty line and other human development indicators to bring the crime rate down.

References

1. NCRB.: Ministry of Home Affairs. Crime in India, 2015- Statistics (2015). <http://ncrb.nic.in>
2. Thompson Reuters.: Canada Best G 20 country to be a Woman; India Worst. <http://in.reuters.com/article/g20-women-idINDEE85C00420120613>
3. Heise, L.L., Pitanguy, J., Germain, A.: Violence against women. In: The Hidden Health Burden. World Bank Discussion paper no. 255 (1994)
4. Indian Express.: NCRB Report: these 6 Indian cities have highest rate of crime against women. <http://indianexpress.com/article/india/india-news-india/ncrb-report-2015-six-indian-cities-most-unsafe-for-women-3005926/>
5. Chitnis, S.: Chapter three in Shrin Kudchekar (eds.) Violence against women and Women against violence, Perfect International Publication, New Delhi (1998)
6. Dutta, N., Rishi, M., Roy, S., Umashankar, V.: Risk factors for domestic violence—an empirical analysis for Indian states. *J. Dev. Areas* **50**(3), 241–259 (2016)
7. Mitra, A., Singh, P.: Human capital attainment and gender empowerment: the Kerala paradox. *Soc. Sci. Q.* **88**(5), 1227–1242 (2007)
8. Land, K.C., McCall, P.L., Cohen, L.E.: Characteristics of US cities with extreme (high or low) crime rates: results of discriminant analyses of 1960, 1970, and 1980 Data. *Soc. Indic. Res.* **24**(3), 209–231 (1991)

9. Williams, K., Ralph, G.: A Multivariate Statistical Analysis of Crime rate in US Cities (2004)
10. Chiricos, Theodore G.: Rates of crime and unemployment - an analysis of aggregate research evidence. *Soc. Probl.* **34**(2), 187–212 (1987)
11. Downs, B.: A critical re-examination of the social and political characteristics of Riot cities. *Soc. Sci. Q.* **15**, 349–360 (1970)
12. Becker, G.: Crime and punishment: an economic approach. *J. Polit. Econ.* **76**(2), 169–217 (1968)
13. Fleisher, B.M.: The effect of income on delinquency. *Am. Econ. Rev.* **56**(1), 118–137 (1966)
14. Singell, C.D.: An examination of the empirical relationship between unemployment and juvenile delinquency. *Am. J. Econ. Sociol.* **26**(4), 377–386 (1967)
15. Statistics on Sex Ratio, for the Year 2011. <http://www.census2011.co.in/sexratio.php>
16. Statistics on Sex Ratio for the Year 2001. <http://www.indiaonlinepages.com/population/sex-ratio-of-india.html>
17. Statistics on Literacy Rate for the Year 2001. http://www.nlm.nic.in/literacy01_nlm.htm and
18. Statistics on Literacy Rate for the Year 2011. <http://www.census2011.co.in/literacy.php>
19. Statistics on Poverty Rate. <http://www.census2011.co.in/literacy.php>
20. Statistics on Net SDP. www.mapsofindia.com/maps/india/percapitaincome.htm
21. Statistics on Human Development Rank. http://www.nipfp.org.in/media/medialibrary/2014/06/WP_2014_139.pdf
22. Statistics on Unemployment rate. https://en.wikipedia.org/wiki/Indian_states_ranked_by_unemployment

Zero Pronouns and Their Resolution in Sanskrit Texts

Madhav Gopal¹ and Girish Nath Jha²

¹ Centre for Linguistics, SLL & CS, Jawaharlal Nehru University,
New Delhi, India

mgopalt@gmail.com

² Special Centre for Sanskrit Studies, Jawaharlal Nehru University,
New Delhi, India

girishjha@gmail.com

Abstract. This paper attempts to formulate an algorithm for resolving Sanskrit null/zero pronouns, specifically null subjects. For this purpose we have conducted an extensive study of two Sanskrit texts, namely, *Panchatantra* and *Hitopadesha* – the *neeti* texts of Sanskrit literature – focusing on null arguments of finite verbs. Use of zero or null pronouns in the language is a massive phenomenon. Null pronouns are empty slots of obligatory arguments of a verb in a finite clause. Sanskrit verbs may take one, two, three or four arguments, depending on their subcategorization. We have observed in the corpus that subject, direct object, indirect object and possessive are the syntactic categories which could be dropped. We have formulated an algorithm using heuristic rules, depending mostly on agreement features of the verb and its subject. We have used the POS annotated data of the above mentioned texts for this study. The whole enterprise involves linguistic analysis of zero pronouns and then determining antecedents of zero subject pronouns after detecting them automatically in the input text.

Keywords: Sanskrit null pronouns · Anaphor resolution · Pro-drop

1 Introduction

Sanskrit is a pro-drop language, that is, some obligatory arguments of verb are omitted in a sentence. We have observed in the Sanskrit corpus that subject, direct object, indirect object and possessive are the syntactic categories which could be omitted. However, in this paper we will offer an algorithm for the subject category only. The language exhibits a strong agreement between subject and verb of an active sentence. Sanskrit verb encodes the person and number features of its subject, and due to which the subject is often dropped, as the information regarding the subject can be sought from the verb itself. In the text of *Panchatantra* (PT) and *Hitopadesha* (HP) it has been well-observed that zero pronouns, i.e. pro-drops, occur frequently. The pro-dropping

phenomenon is a problem for the activity of translation, especially while rendering Sanskrit texts into English or English like languages which do not allow argument omission. When pro-drop constructions are translated into other languages these pros/zero/null pronouns have to be tracked down, because the grammatical system of the target language may not allow this pro-dropping. For intelligence possessing humans this resolution is not a difficult task, as they can comfortably recover the omitted arguments from the context. But if a machine has to translate Sanskrit texts into English like languages, then it will face enormous difficulty as it does not have that intelligence to recover what is not given in the sentence. To solve this problem of machine we have conducted this research and finalized an algorithm to help machine recover the omitted arguments. For the resolution of these dropped arguments, the algorithm relies entirely upon the verb that will give it person and number features of the dropped subject. The texts of PT and HP were digitized and POS tagged with Indic Language POS Tagset (IL-POST) developed by Microsoft Research India (Jha et al. 2009). The IL-POST annotates morphosyntactic features also along with grammatical categories.

2 Methodology

We have first done sandhi (external) splitting of the Sanskrit texts mentioned above, that were digitized in the UTF-8 format. Then with the help of IL-POSTS (Jha et al. 2009) we have annotated the corpus with POS tags. In this annotation scheme tags are hierarchically organized. Along with main grammatical categories, their sub-categories are also specified and the morpho-syntactic features such as person, number, gender, tense, active, passive, case etc. are also marked with the tag. These attributes are instrumental in detecting the null subject and also in the resolution of its antecedent. The tagged corpus was useful for this research only because of annotation of these very features. After an extensive study of the texts we have formulated an algorithm to resolve null subjects which is reported in the Sect. 5 and have implemented the system using java platform. The results have been analysed in the Sect. 7.

3 Sanskrit Zero Pronouns

Sanskrit has a well developed system of pronouns. Many studies like Speijer (1886), Whitney (1889), Kale (1995) and Gopal (2012) have discussed overt pronouns at length but the covert pronouns or pro-drops are not dealt with in these works. Sanskrit allows pro-dropping frequently. Like many other South Asian languages, any obligatory argument of verb can be dropped - be it subject, direct object or indirect object. There is a strong agreement between the verb and its subject in a sentence. Sanskrit verb encodes the person and number features of its subject, and this is the reason that

subject is more often dropped than any other argument, as the information regarding the subject can be recovered from the verb itself. Unlike Hindi verbs, Sanskrit verbs do not agree with the objects. Moreover, grammatical subjects of passive clauses which are logical objects of their active counterparts are also inclined to be omitted. In certain situations possessor is also dropped.

3.1 Classification of Sanskrit Null Pronouns

Primarily, zero pronouns could be classified into two categories: anaphoric and non-anaphoric. Those zero pronouns are anaphoric which have antecedents in the preceding or following discourse and those which do not have antecedents and serve as a source of generic interpretation are called non-anaphoric. To study the null pronouns of the Korean language Han (2006) has categorized them in the following categories.

Table 1. Classification of Korean zero pronouns by Han (2006)

Text dependent use	Anaphoric zero pronoun
	Discourse-deictic zero pronoun
Text independent use	Deictic zero pronoun
	Indefinite personal zero pronoun
	General situational zero pronoun

This classification is very insightful for classifying Sanskrit null pronouns. We have adapted the Table 1 for Sanskrit null pronouns with minor changes as per the requirement of Sanskrit data. For Sanskrit, we have the following Table 2:

Table 2. Classification of Sanskrit zero pronouns

Text dependent use	Anaphoric null pronoun
	Discourse-deictic null pronoun
Text independent use	Deictic null pronoun
	Generic null pronoun

3.1.1 Anaphoric Null Pronouns

In Sanskrit, as we have said above, anaphoric null pronouns occur in a variety of syntactic contexts: subject, direct object, indirect object, and possessor. Null subjects of finite clauses are mostly unrealized as they can be easily recovered by the agreement features of the verb. The topic element of linguistic utterances is usually expressed in nominative case. Once it is introduced in the discourse and activated cognitively in the mind of the hearer, it is usually dropped, as illustrated in (1).

- (1) u1. तत्र च लघुपतनकः नाम वायसः प्रतिवसति स्म ।
tatra ca laghupataṇakaḥḥ nama vayasah
 there and Laghupataṇaka.NOM.MAS.SG named crow.NOM.MAS.SG
pratiwasati sma.
 live.PRS.3.SG was
 ‘There was living a crow named Laghupataṇaka.’
- u2. कदाचित् प्राणयात्रार्थम् पुरम् उद्दिश्य प्रचलितः यावत् ∅ पश्यति,
kadacit praṇayatrartham puram uddishya pracalitaḥ yawat ∅
 sometime for food city.ACC aimed go.PSPL when (SBJ)
pashyati,
 see.PRS.3SG
 ‘Once upon a time, going towards the city for food, when (he) looks,’
- u3. तावत् जालहस्तः अतिकृष्णतनुः, स्फुटितचरणः,
tāwat jālahastah ati-kṛṣṇa-tanuḥ, sphuṭita-caraṇah,
 then net-hand.SG.NOM very-black-body torn-foot-M.SG.NOM
 ऊर्ध्वकेशः यमकिङ्कराकारः नरः ∅
ūrdhwakeshaḥ yamakinkarākārah narah ∅
 vertical-hair.NOM Yama-servant-appearance.M.NOM man.M.NOM (his)
 संमुखः बभूव ।
sammukhaḥ babhūwa.
 in front of be.PST.3SG
 ‘Then, a man with net in his hand, with a very black body, big feet, flying hair, looking like a servant of god Yama, came in front of him’.
- u4. अथ तम् दृष्ट्वा शङ्कितमना ∅ व्यचिन्तयत् - यत्
atha tam dṛṣṭvā shankitamanaḥ ∅ vyacintayat – yat
 after him seeing doubtfully (SBJ) thought - that
 after seeing him, (he) thought doubtfully, - that

¹In this example the topic Laghupataṇaka is introduced in u1 as the subject of the initial sentence of this discourse segment. In u2 and u4 subject arguments are dropped. These pro-drops are co-referential with the subject of u1. In u3 a possessive has also been elided which is co-referential with the same entity. In the example (2) below, direct object is omitted which is coreferential with Bhasurak.

¹ The data in this paper is presented uniformly. Every discourse segment is numbered and the utterances within the segments are further numbered as u1, u2, u3...un. The assumed null forms are indicated by the sign ∅ at their most natural place in the clause. Sanskrit examples are taken from *Panchatantram* and their translations are literal.

- (2) अहम् भासुरकम् प्रकोप्य
aham bhasurkam prakopya
 1SG.NOM Bhasuraka.M.ACC anger.PCPL.INS.SG
 स्वबुद्ध्या Ø अस्मिन् कूपे पातयिष्यामि ।”
swabuddhya Ø asmin kūpe patayiṣyami.”
 self-intelligence.F.INS (DOBJ) this.M.LOC well.M.SG.LOC demolish.1SG.FUT
 ‘After getting Bhasurak angered, I will throw (him) in the well.’

Following is an example showing omission of an indirect object.

- (3) दुर्भिक्षत्वात् जनः बुभुक्षापीडितः कोपि
durbhikṣa-twāt jaṇaḥ bubhukṣāpeeditaḥ kopi
 drought-reason.ABL person hunger-suffering-sg.nom one-EMPH
 बलिमात्रम् अपि न Ø प्रयच्छति ।
balimātram api na Ø prayacchati.
 Sacrifice.SG.ACC EMPH NEG (IOBJ) give.PRS.3SG
 ‘Due to drought, any hungry person is not giving even sacrifice (to any one).’

Though it is very rare, it is also possible in the language to drop all the required arguments of a verb. The causative form of the verb *drishir* ‘to see’ requires a subject, a direct object and an indirect object but all of them are simultaneously omitted in the utterance segment (4) u3.

- (4) u1. त्वाम् दृष्ट्वा दूरतः अपि चौरसिंहः प्रविष्टः स्वम् दुर्गम् ।
tvām drṣṭvā dūrataḥ api caurasimhaḥ praviṣṭaḥ swam durgam
 2SG.ACC seeing remotely EMPH thief-lion entered self.ACC fort.ACC
 Seeing you remotely, the thief lion has entered in his fort.
- u2. तत् आगच्छ,
tat āgaccha,
 then come.IMP.2SG
 ‘Then come.’
- u3. येन Ø Ø Ø दर्शयामि इति ।
yena Ø Ø Ø darshayāmi iti
 so_that (SUB DOBJ IOBJ) show.PRS.1SG QUOT
 ‘So that, (I) show (the lion) (to you).’

In the corpus we have noted that speaker and/or addressee of an utterance are frequently dropped, as is shown below:

- (5) u1. दमनकः आह-
damaṇakaḥ āḥa-
 Damanak.SG.NOM say.PRS.3SG
 Damaṇaka says -
- u2. “किम् स्वामिनम् पिङ्गलकम् अपि न जानासि ?”
kim swāminam pingalakam api na jāñāsi?
 What master.M.SG.ACC Pingalaka.M.SG.ACC EMPH NEG know.2SG.PRS
 ‘(You) don’t know even master Pingalaka?’
- u3. तत् Ø क्षणम् प्रतिपालय ।
tat Ø kṣaṇam pratipālaya.
 Then (SBJ) moment.ACC wait.2SG.IMP
 ‘Then (you) wait a moment.’
- u4. Ø फलेन एव Ø ज्ञास्यसि ।
Ø phalena eva Ø jñāsyasi
 (SBJ) Fruit.N.SG.INS EMPH (OBJ) know.2SG.FUT
 (You) will know (him) by fruit.
- u5. देवशर्मा आह-
Dewasharmā āḥa-
 Dewasharma.SG.NOM say.PRS.3SG
 Dewasharma said-
- u6. “वत्स ! Ø अनुग्रहम् ते करिष्यामि ।
watsa! Ø anugraḥam te kariṣyāmi.
 Son! (SBJ) Grace.ACC 2SG.DAT do.1SG.FUT
 ‘Son! (I) will do good to you.’

The use of zero pronouns in Sanskrit is not only anaphoric but cataphoric also. There are limited usages of such expressions in the texts. In the following sentence the subject of the verb *karoti* ‘does’ is to be supplied from the succeeding clause:

- (6) u1 करटकः आह-
karatakaḥ āḥa-
 Kartataka.SG.NOM say.PRS.3SG
 Karataka says-
- u2 “यद्यपि त्वदीयवचनम् न करोति,
yadyapi tvadeeya-vacaṇam na karoti,
 though 2POSS.SG-word-N.SG.ACC NEG does,
 ‘Though (he=the master) does not pay any heed to your words,’
- u3 तथापि स्वामि स्वदोषनाशाय वाच्यः ।
tathāpi swāmi sva-doṣa-nāsha-āya vācyaḥ.
 still master self-fault-destuction-SG.DAT speakable
 ‘Still for the destruction of your own faults you should speak to him.’

3.1.2 Discourse-Deictic Null Pronoun

Sometimes in a discourse, propositions, situations or events also get pronominalised. Such pronominal forms are called discourse deictic pronouns. Empty forms of such pronouns have also been sparsely used in the texts. In the following utterance (7) u1*etat* is an example of textual-deictic pronoun and in (7) u4 the same entity is dropped making it an instance of textual-deictic null pronoun.

7. u1 एतत् उक्त्वा सः भूयः अपि प्राह –

etat uktvā saḥ bhūyaḥ api prāḥ-

this saying 3SG.M.NOM again EMPH say.PRS.3SG

Having said this, he further says-

u2 “अथ ज्ञायते तस्य क्रमणमार्गः ?”

atha jñayate tasya kramaṇa-mārgaḥ?

After know.3SG.PAS his arrival.path.SG.NOM

‘Do you know his path of arrival?’

u3 ताम्रचूडः आह –

tāmracūdaḥ āḥa-

Tamrachood.M.NOM say.PRS.3SG

Tamracuda says-

u4 “भगवन्! Ø ज्ञायते,

Bhadawan Ø jñāyate,

Lord.SG.VOC (SBJ) know.3SG.PRS.PAS

‘Lord! (that) is known to me.’

u5 यतः Ø एकाकी न समागच्छति ।

yataḥ Ø ekāki na samāgacchati.

Because (SBJ) alone NEG come.3SG.PRS

‘Because (he) does not come alone.’

3.1.3 Deictic Zero Pronouns

Deictic pronouns are linguistic entities which refer directly to the personal, temporal or locational characteristics of the situation within which an utterance takes place, whose meaning is thus relative to that situation. Their referent could be interlocutors of the discourse, an object or a third person relativised to that situation. In the (8) A below, the

king asks a question using an overt deictic but in the answer in (B) these required deictics are missing which have been supplied in the corresponding English translation.

(8) A. राजा आह –

rājā āha-
king.M.NOM say.PRS.3SG
'The king says-'

कः अयम् ?

kaḥ ayam
who this.M.NOM
'Who is this?'

कुतः \emptyset समायातः ?

kutaḥ \emptyset samāyātaḥ
from where (SBJ) came.M.SG
'Where has (this) come from?'

B. ते ऊचुः –

te ūcuḥ-
3PL.M.DST say.PL.PST
They said-

\emptyset हिरण्यगर्भनाम्नः राजहंसस्य अनुचरः, \emptyset कर्पूरद्वीपात् आगतः ।

\emptyset hiraṇyagarbhanāmnāḥ rājahansasya anucarāḥ
(this) Hiraṇyagarbha-named Rajahansa-gen.sg servant.m.sg.nom

\emptyset karpūradweepāt āgataḥ.
(this) Karpoordeepa-ABL.SG come.PSPL.M.SG

'(This) is the servant of the king Hiraṇyagarbha Rajahansa, (this) has come from Karpooradweepa.'

3.1.4 Generic Null Pronouns

Like Korean, Sanskrit too employs generic null pronouns which do not refer a particular individual but people in general. In (9) u1 indirect object of the verb *dā* 'give' is elided whereas subject and direct object are overtly present. Since this is a statement regarding people in general, no specific individual is named. In (9) u2 utterance indefinite generic subject (impersonal subject) is dropped, for the same reason.

(9) u1. विद्या ∅ ददाति विनयम्

vidyā ∅ *dadāti* *vinayam*
 learning.F.SG.NOM (IOBJ) give-PRS.3SG politeness.SG.ACC
 ‘Learning gives (one) politeness.’

u2 विनयात् ∅ याति पात्रताम् ।

vinayāt ∅ *yāti* *pātratām*.
 politeness.F.SG.ABL (SBJ) go-PRS.3SG eligibility.F.ACC
 ‘From politeness (one) goes to eligibility.’

4 Detection of Null Pronouns

Being invisible entity the null pronouns require first and foremost their detection. Sanskrit being a “free word order” language, words are ordered quite freely. So, exactly which place a pronoun should occur is not an important question. But it is expected to occur in a sentence, no matter overt or covert. On the basis of the subcategorization of the verb one can decide whether or not there exists any null category. Generally, all Sanskrit verbs require a subject – grammatical or logical. Verbs are subcategorized as intransitive, transitive and ditransitive requiring one, two and three arguments respectively. Though, our corpus does not have any tag for these features, in future research we could incorporate these features in the morpho-syntactic tagging of the corpus. The corpus does not have tags for null entities. So, for detecting null subjects, we look at the verb – its voice, and person, number features. If there is a corresponding nominal/pronominal in the clause then verb has its subject overtly, otherwise covertly. The covert subjects have to be checked in the previous clauses. Usually three previous clauses are sufficient for searching the antecedent of this null entity. To detect arguments other than subject, we would need more rigorous tagging, which we plan to do in future.

5 Algorithm for Resolving Subject Null Pronouns

1. Tokenize each sentence (S) of the input text (POS tagged Sanskrit text), based on a *danḍa*.
2. Tokenize each clause (C) within the tokenized sentences, based on the tag CSB and the punctuation (–) dash.
3. (i) Pick up the C in which V.act or V.pas tag is found anywhere.
 (ii) Or Pick up the C in which KDP (participle) or KDG (gerund) tags occur at the final position of the C.

(iii) Or Pick up the C in which KDP or KDG tags occur at the penultimate position of the C followed by the conjunction च्.

For Null Subject Resolution in Active Voice

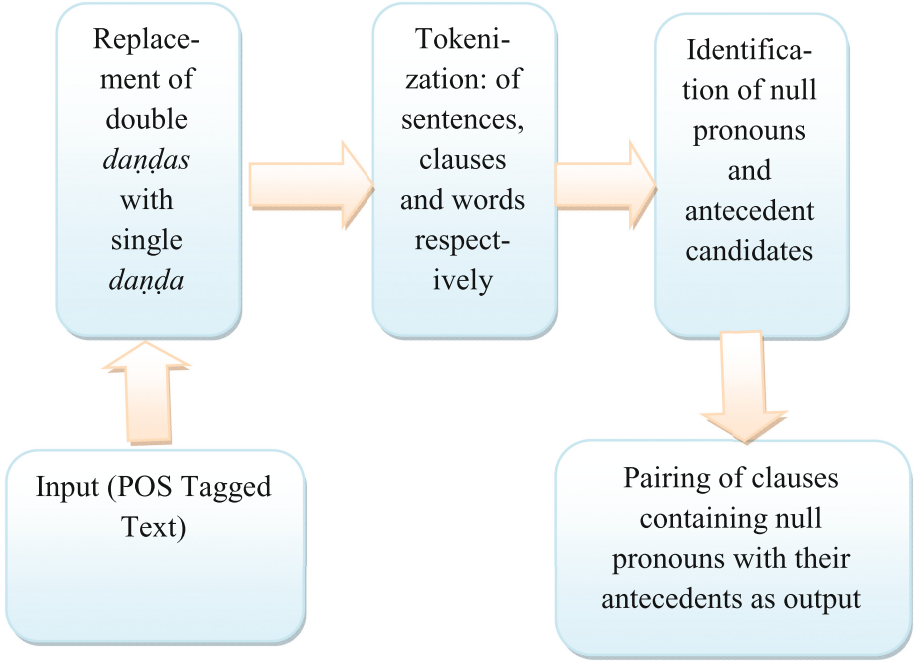
4. If the C contains V.act tag then find in the same C an NP.nom/NC.nom/PPR.nom having same **number** and **person** features as the verb. If found then leave the sentence, as this is not an instance of null subject. If not found then lūk for the same in the previous **three** clauses. A candidate immediately preceding नाम particle will get preference over other candidates. If नाम is not there then the nearest candidate will be the resolution.
- 4.a If V.act.du tag occurs in C then find an NP.nom/NC.nom/PPR.nom with.du. attribute or two instances of an NP.nom/NC.nom/PPR.nom with.sg. attribute. If found then leave the sentence. If not found then look for the same in the previous **three** clauses. The nearest candidate is the antecedent.
- 4.b If V.act.pl tag occurs in C then find an NP.nom/NC.nom/PPR.nom with.pl. attribute or at least three instances of an NP.nom/NC.nom/PPR.nom with.sg. attribute. If found then leave the sentence. If not found then lūk for the same in the previous **three** clauses. The nearest candidate is the antecedent.
5. For all the conditions stipulated in 4, 4a. and 4b. treat KDP and KDG as V.
6. If a C has no NP.nom/NC.nom/PPR.nom but has V.act. occurring at the final position of the C is उवाच, प्रोवाच, ऊचुः, प्रोचुः, ऊचे, आह or प्राह, then look for an NP.nom/NC.nom/PPR.nom tag having same **number** and **person** features of the above mentioned verbs in the previous three clauses. If the immediate utterance is within inverted commas, then go back to the immediate clause preceding the whole quoted discourse segment. The NP.nom/NC.nom/PPR.nom with compatible features will be the antecedent.

For Logical Null Subject Resolution in Passive Voice

7. If the C contains V.pas or KDP or यम्\KDG or ये\KDG or यानि\KDG or यः\KDG or यै\KDG or या\KDG or या\KDG tag then find an NP.ins/NC.ins/PPR.ins tag having any number and person features. If found then leave the sentence.
8. If NP.ins/NC.ins/PPR.ins is not found then look for them in the previous **two** clauses. The nearest candidate is the antecedent. (A candidate preceding *nāma* particle will get preference over other candidates. If *nāma* is not there then the nearest candidate will be the resolution.)

6 System Architecture

The NARSS (an acronym for Null Anaphor Resolution System for Sanskrit) has four modules to undergo for processing before the desired resolution. These are preprocessor which replaces the double *danḍas* with single *danḍas* which works as a delimiter for tokenisation; String tokeniser which tokenises sentences/clauses and later on words; Tag checker which checks the categories of words and consequently determines the presence or absence of null arguments, and finally Morph-checker (morpho-syntactic features checker) to fix the resolution of null anaphors.



Model Diagram of NARSS

6.1 Brief Description of the Processes

The system takes POS annotated data of Sanskrit in UTF-8 format as input. It goes through five processes before it gets the final result as is shown in the diagram. The input is pasted in the text area of the system which is available in the homepage of NARSS, and then it is sent for processing and resolution. The pre-processor of the system replaces the double *daṇḍas* with single *daṇḍas* (it is customary in Sanskrit to use double *daṇḍas* at the end of a verse). Afterwards, the data is tokenized in single sentences delimited by a single *daṇḍa*. After sentence tokenization, within sentences each clause is tokenized with the help of connective tags. And then each word of a clause is tokenized. The system considers a single tokenised clause at a time for resolving null anaphors. In each tokenized clause the system first checks whether it contains or not a verb, a participle or a gerund. If they are not in that clause then it leaves that clause as our system is based on agreement features of verb and its subject. And verb being absent from the clause, there could not be established any agreement. The system, then, considers next clause for the same. And if a verb, a participle or a gerund exists in the clause, it searches for a corresponding subject on the basis of morphosyntactic features as per our rules (see the algorithm in the Sect. 5). If a subject

entity is found in a clause then that clause too is abandoned as it does not have a null subject. And if the subject is not found then the system holds the clause and searches the antecedent of the null subject. The resolution of this antecedent for the null subject is done as per our algorithm. After resolving the null argument in a sentence the system moves on to the next sentence and operates likewise till the end of the string. The candidates for being the antecedents of null subjects are common nouns (NC), proper nouns (NP), personal pronouns (PPR) and relative pronouns (PRL) as per our generalisations in the PT and HP. The antecedents of PPR and PRL will be resolved further. All these categories are simply identified on the basis of their morphosyntactic tags. Thus POS tagging information is very crucial for anaphora resolution in this system.

7 Result Analysis

As we have noted above, Sanskrit allows not only subjects to drop but also direct objects, indirect objects, and possessors. In this experiment we were dealing with the resolution of null subjects only. Following is a tabular description of our tentative results. We are still in the process of improving our results (Table 3).

Table 3. Evaluation results of Sanskrit zero pronouns

Voice	Clauses with explicit verb/participle	Clauses with null subject	Resolution	
			Correct	Wrong
Active	4860	1824	1122	702
Passive	4456	1240	934	306
Total	9316	3064	2056	1008

8 Conclusion and Future Research

In this paper we have presented an ongoing work on Sanskrit null pronouns based on two story texts, namely, *Panchantra* and *Hitopadesha*. A classification of null pronouns was done according to the linguistic behavior of the pronouns concerned. We have taken examples from the said texts only. We have formalized a scheme of tokenizing different types of clauses as per our convenience of dissecting discourse utterances. Each tokenized clause has a finite verb or participles. On the basis of agreement features of Verb and its subject, an algorithm has been evolved and presented. Using this algorithm we have implemented the system NARSS which is giving us encouraging results. We plan to put this system online (at <http://sanskrit.jnu.ac.in>) for resolving Sanskrit null pronouns developed in the Java platform. The system will be enhanced by incorporating rules for resolving object, indirect object and possessor categories of null construction.

Appendix: A Sample of Tagged Text We Used for This Research

(\PU नृपसेवकवानरकथा\NP.fem.sg.nom.i)\PU कस्यचित्\CX राज्ञः\NC.mas.sg.gen.vi नित्यम्\CAD
 वानरः\NC.mas.sg.nom.i अतिभक्तिपरः\JJ.mas.sg.nom.i.n.n अङ्गसेवकः\JJ.mas.sg.nom.i.n.n
 अन्तःपुरे\NC.neu.sg.loc.vii अपि\CAD प्रतिषिद्धप्रसरः\JJ.mas.sg.nom.i.n.n
 अतिविश्वासस्थानम्\JJ.neu.sg.nom.i.n.n अभूत्\V.act.sg.3.aor.n \PU एकदा\CAD
 राज्ञः\NC.mas.sg.gen.vi निद्रागतस्य\JJ.mas.sg.gen.vi.n.n वानरे\NC.mas.sg.loc.vii
 व्यजनम्\NC.neu.sg.acc.ii नीत्वा\CGD वायुम्\NC.mas.sg.acc.ii विदधति\KDP.mas.sg.loc.vii
 राज्ञः\NC.mas.sg.gen.vi वक्षःस्थलस्य\NC.neu.sg.gen.vi उपरि\CPP मक्षिका\NC.fem.sg.0.i
 उपविष्टा\KDP.fem.sg.0.i \PU व्यजनेन\NC.neu.sg.ins.iii मुहुः\CAD मुहुः\CRD
 निषिध्यमाना\KDP.fem.sg.nom.i अपि\CAD पुनः\CAD पुनः\CRD तत्र\CAD एव\CEM
 उपविशति\V.act.sg.3.prs.n \PU

References

- Gopal, M.: Anaphor Resolution in the Sanskrit Text Panchatantra. LAP LAMBERT Academic Publishing, Saarbrucken (2012)
- Jha, G.N., Gopal, M., Mishra, D.: Annotating Sanskrit Corpus: adapting IL-POSTS. In: Vetulani, Z. (ed.) Lecture Notes in Artificial Intelligence, pp. 371–379. Springer, Heidelberg (2009). ISSN 0302-9743, ISBN 978-3-642-20094-6
- Han, N.-R.: Korean zero pronouns: analysis and resolution. Ph.D. thesis, Department of Linguistics, University of Pennsylvania (2006)
- Kale, M.R.: A Higher Sanskrit Grammar. MLBD Publishers, New Delhi (1995)
- Speijer, J.S.: Sanskrit Syntax. Motilal Banarsidass Pvt. Ltd., New Delhi (1886). Repr. (2006)
- Whitney, W.D.: Sanskrit Grammar. Munshiram Manoharlal Publishers Pvt. Ltd., New Delhi (1889). Repr. 2004
- <http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2011T04>. Accessed 24 May 2017)

Semantic Analysis Using Pairwise Sentence Comparison with Word Embeddings

Vijay Krishna Menon^(✉), Sabdhi M., Harikumar K., and Soman K.P.

Center for Computational Engineering and Networking, Amrita University,
Coimbatore, India

m.vijaykrishna@cb.amrita.edu

Abstract. Comparing the semantics of a pair of sentences has been an interesting yet unstructured problem. Semantic analysis is mostly elusive due to the fact that the semantics of Natural language constructs cannot be measured, let alone be compared to one another. Methods like Latent Semantic Analysis(LSA) and Latent Dichlaret Analysis(LDA) are able to capture broader semantics between documents, but their contribution in pairwise comparison tasks which require deeper semantics may be limited. In this paper we present a local alignment based scoring scheme for sentence pairs using word embeddings and how this can be used as a feature for some popular text analysis tasks such as summarization, paraphrase comparison, topic profiling and other semantic comparison tasks. We also present a theoretical analysis on the metrics used in this approach and a separability argument using t-SNE plots. Furthermore we detail our Spark implementation model for the pairwise comparison and summarization.

Keywords: Pairwise comparison · Semantic alignment · Smith-Waterman · Word embeddings · Apache Spark · Word vectors · Text summarization

1 Overview

Text similarity is a well researched domain in which semantic similarity forms a much smaller subset task [12]. While text is easy to align and score thus quantifying the measure of similarity between two textual strings, finding a deeper match between the real information represented by the query and template text phrases is an elusive problem. To be precise, comparing the inner meanings rather than plain strings, is difficult. There are numerous text similarity measures and algorithms that can be used as given by [6]. We will use Smith-Waterman local alignment for comparing sentence pairs; pairwise comparisons can be thus quantified [11]. This can be treated as a distance measure or metric of relatedness between the two sentences. The main issue in such a method is the complexity of computing the pairwise distance matrix which will be used for further more

specific tasks of interest. That being said, these computations can be easily parallelised as there are no dependencies once the pairs are identified. For this, we employ the popular Apache Spark[®] framework to deploy the computations to a cluster.

Most recent developments in text analysis is the usage of deep learning as the basic technique to yield, over state of the art results. Deep learning can be employed to do pairwise phrase semantics comparison using deep neural architectures from scratch [9]. The main issue with deep learning is the computing requirements for training these massive neural nets. Our approach to this was to use *pre-trained word embeddings* so that the training part can be foregone; pre-trained word vectors are available in plenty and so are the tools used for generating these vectors given a monolingual corpus, such as the *word2vec* tool by Google[®]. For using an alignment algorithm, we required a scoring scheme or a score matrix which in many cases is statistically learned or estimated. But when it comes to words and semantics we have seen that this does not yield significant comparisons since representations such as one hot vectors as used in LSA and the n-gram vectors used for other NLP tasks, has not yielded any significant improvements in semantic tasks over this time since they have been introduced. The other option will be to construct a linguistic scoring model that will match word similarities based on POS tags, synonyms, co-locations, tense etc. This entails considerable linguistic work and will be extremely language specific. Word embedding on the other hand are semantics preserving [16] in nature, contributing to most linguistic aspects required for a good semantic comparison. The idea is simple enough; we use a distance between the corresponding word vectors of the words as a score, indicating their semantic similarity. We have chosen the GloVe [17] vectors trained on Wikipedia 2014 corpus with six billion tokens; each word vector $w \in \mathbb{R}^{300}$.

2 Overlap Finding Methodology

Given a pair of sentences, we have to find the semantic overlap between them. The best way to do this will be to do a substring alignment of the pair and trace back the local matches that have more than a significant trace score. This kind of alignment is also called *local alignment* and is used in comparative bioinformatics for aligning protein or gene sequences. We employ the exact same algorithm proposed by Temple F Smith and Michael S Waterman [20] and modified by Robert Irving [11] for use in text overlap detection. Irving gave a cut-off modification that will slice out all overlapping alignments, so the match will not be counted more than once. Even though the Smith-Waterman algorithm is designed for character level alignment, we can use it to align sentences at word level. Smith-Waterman local alignment is a dynamic programming algorithm, that will align all matching sub-sequences in the given pair of sentences. Let us say that s_t and s_q are 2 sentences with length t and q . The alignment matrix, S_{tq} will be of dimensions $(t + 1) \times (q + 1)$, where the first row and column is for *gap*

alignment. The S_{tq} matrix is populated using the standard recurrence relation as given below.

$$S_{tq}(i, j) = \max \begin{cases} 0, \\ S_{tq}(i - 1, j - 1) + \delta(s_t(i), s_q(j)), \\ S_{tq}(i, j - 1) - G_p, \\ S_{tq}(i - 1, j) - G_p \end{cases} \tag{1}$$

$$\forall_i \in [0, t], \forall_j \in [0, q]$$

$$S_{tq}(i, 0) = 0, \forall i$$

$$S_{tq}(0, j) = 0, \forall j$$

Here δ is a scoring function for the i^{th} word in s_t and the j^{th} word in s_q . However this formulation is insufficient if we want to find semantic similarity of sentence pairs as it will *align same sub-sequence to multiple matching sub-sequences of words* which will be problematic since we add the scores of all local alignments (multiple local alignments might occur with in each sentence pair) that exists. This is called the *overlapping local alignments* problem. To rectify this issue we use the afore mentioned cut-off modification in [11] which we shall not discuss here in detail.

The next important aspect is measurement of the overlap of the sentence pairs. As mentioned above, the alignment process is guided by an adept scoring scheme that reflects our objective for alignment. Here, it is semantic similarity of the two sentences which contain words. So, Like in Fig. 1, we score each column of each local alignment. *Similarity* is just the cumulative score of all non

I	went	----	----	home	then
I	came	to	your	house	yesterday

Fig. 1. A simple sentence alignment model where words are scored according to their relatedness.

	I was ----a	similarity :4.1
	I was little	
	the ----war --brokeout	similarity :6.1
	The fighting took-place	
was -----a	-----boy when the war brokeout	similarity :2.3
The fighting took-place	when --I was --little	
Total Similarity		0.92

Fig. 2. Several local alignments of the sentences *I was a boy when the war brokeout* and *The fighting took-place when I was little*. The score of all alignments contribute to the semantic similarity between these sentences

overlapping local alignments between the pair, that scores above a set threshold [11]. Figure 2 shows how multiple local alignments can be obtained from a pair of sentences. The scoring scheme used here assigns high scores to alignment of words with apparently the same meaning like **war** with **fighting** and **boy** with **little** which has been manually annotated. This is an ideal scoring scheme, but

Table 1. Readings obtained using four similarity metrics for text summarization using pairwise analysis. Only some performance related observations are tabulated here; no comparisons are given. The sentences in the summary are not affected by the distance metric chosen. But execution times differ based on the difficulty in computing these distances.

Sentences	Similarity type	Time (min)	Density %	Summary count
102	Cosine-Distance	5	8	17
	Euclidean-Distance	6	49	16
	Standard-Correlation	5	8	17
	Pearson-Correlation	5	8	17
134	Cosine-Distance	6	9	24
	Euclidean-Distance	7	49	22
	Standard-Correlation	8	9	24
	Pearson-Correlation	7	8	24
169	Cosine-Distance	8	8	32
	Euclidean-Distance	4	52	28
	Standard-Correlation	6	8	32
	Person-Correlation	5	8	31
194	Cosine-Distance	9	8	33
	Euclidean-Distance	6	48	30
	Standard-Correlation	8	8	33
	Pearson-Correlation	10	7	33
216	Cosine-Distance	10	7	37
	Euclidean-Distance	6	49	35
	Standard-Correlation	9	7	37
	Pearson-Correlation	13	7	36
226	Cosine-Distance	12	7	36
	Euclidean-Distance	7	47	42
	Standard-Correlation	12	7	36
	Pearson-Correlation	14	7	36
250	Cosine-Distance	15	6	39
	Euclidean-Distance	6	44	46
	Standard-Correlation	10	6	39
	Pearson-Correlation	15	6	39

it is impossible to define this for the entire language. We have found that distributed representations of words [16], popularly called as word embeddings, are semantically preserving, i.e., they are able to learn relational semantics between words and translate this in to spatial positions. These word representations are also called as word vectors, since they represent a unique position vector for each word occurring in the training corpus. The vectors are obtained by learning word co-locations in a monolingual text corpus, by regressing them through a deep neural net, that translates simple n-grams to some m-dimensional vectors. To understand more on how word vectors are trained please refer the works of Ronan Collobert, Jason Weston [4], Joseph Turian, Yoshua Bengio [21], Tomas Mikolov [16] and Richard Socher [17]. Word vectors give us the leverage in comparing words within alignments. A simple *cosine distance* measure can act as a similarity value. We use this as a score between the words while aligning sentences. Cosine similarity is the most common place metric. However we can also use standard correlation and Pearson’s correlation both of which has the same range (value in $[-1,1]$). Euclidean distance can also be employed, given the vectors are normalised with in the unit sphere. We have used multiple distance measures for scoring, which we have presented in Table 1 with their performance observations, which shall be detailed later.

3 Pairwise Matrix and the Sentence Vector

Given a discourse text with m sentences (without the same sentence repeating), A pairwise distance matrix, D_p can be computed by aligning the pairs of all sentences. Each sentence s_i with $i \in [0, m - 1]$ is represented by i^{th} column in D_p . This will represent the sentence as a m dimensional vector. The column space of this matrix represents the topic subspace of the entire discourse. It is possible to observe the various subtopic clusters within, but since we deal with texts having many sentences, visualizing this phenomenon is impossible as we deal with vector spaces far beyond 3 dimensions. We were able to generate some amicable plots using t-Distributed Stochastic Neighbourhood Embedding (t-SNE) [14] algorithm and mapping the sentence vectors to 2-dimensions. Figure 3 represents the 2 dimensional t-SNE plot of sentences (the actual vectors are high dimensional) from samples of Wikipedia from multiple articles. This figure shows clusters of sentences are formed and they are very separable using the pairwise feature vectors. The pairwise distance matrix D_p (further referred as the pair matrix) is a symmetric (positive definite) matrix which is extremely sparse, as it should be. The computation of this matrix is done only for the lower triangular values and then reconstructed to form the full matrix; this simple tweak will not change the time complexity of the algorithm but it can make execution twice as fast. The values are all normalised between zero and one, so that it can be treated like a probability of semantic match. Consider the sentence pairs (s_t, s_q)

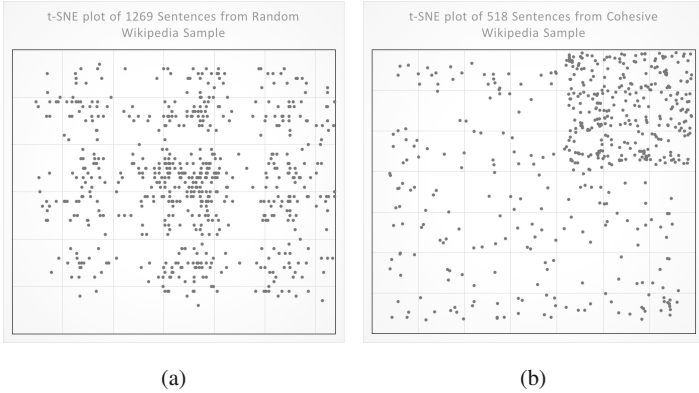


Fig. 3. t-SNE plot showing clustering of subtopics. The scatter plot (a) has 1269 sentences from random Wikipedia sample text. Multiple separate clusters are clearly observable as many topics categories are captured. The one in (b) is plotted with 518 sentences form cohesive Wikipedia sample text with a dominant topic area. Here we can see one major cluster and many loose points. The pairwise vectors will have the dimensionality equal to the number of sentences

(template, query), the local alignment a_i^{tq} is the i^{th} alignment and $\Delta(a_i^{tq})$ is its total score, then the similarity is computed as follows.

$$S(s_t, s_q) = \begin{cases} 1 - \frac{1}{\alpha} & \alpha \neq 0 \\ 0 & otherwise \end{cases} \quad \alpha = \sum_{i=0}^K \Delta(a_i^{tq}) \quad (2)$$

An alternative approach to this will be to take the maximal alignment instead of sum of all alignments. This approach can indeed save some computation as Irving’s cut-off need not be used in it. But it might miss out on vital matches in longer sentences. The main motivation to go for the summation approach was to use this in general domain text analysis. The sentences in general domain can be longer and usually have several topics referred in it. This might lead to some information loss if we used only the maximal scoring alignment and shunned all the smaller ones. On the contrary, specific topic related text perform very well in such a method since stray topic semantics are minimal. We advice the use of both technique for similarity computation depending upon the data in hand. We can also reverse this measure to find dissimilarity between sentences by simply calculating $\frac{1}{\alpha}$.

4 A Rough Extractive Summarization

We have used the pairwise matrix to localise the dominant topic subspace. As mentioned before the columns of the pair matrix spans the topic subspace of the text in an m-dimensional vector space. Since it is very sparse with many empty columns as we have observed, despite being a square matrix it cannot be full

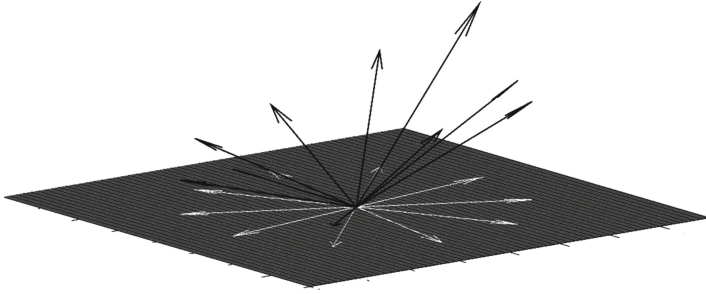


Fig. 4. Visualisation of dominant topic subspace and projections. The white shadows are projections of the black sentence vectors and the black plane is the dominant topic subspace. The summarization strategy is to extract the sentence that create significant length white projection on the plane.

rank. Our objective is to find the dominant topic subspace which contributes to the majority of the sense conveyed by the text. We call this high content summary. We do either an Eigen decomposition of it or simply apply PCA on it to find the main principal vectors spanning the dominant topic subspace. Let us consider D_p as our pair matrix. E_{sub} will be the principal component matrix obtained by applying PCA on D_p . Now when we multiply both of these matrices we obtain the projection matrix P . The columns of the P consists of the projection vectors of the original sentence vectors onto the dominant topic subspace as depicted in Fig. 4. From Eq. 2, we know S_{tq} is the similarity measure between s_t and s_q pair of sentences, then:

$$D_p = \begin{bmatrix} S_{00} & S_{01} & \dots & S_{0m} \\ S_{10} & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ S_{m0} & \cdot & \cdot & S_{mm} \end{bmatrix} \tag{3}$$

Taking SVD on D_p , we get

$$D_p = U \Sigma V^T \tag{4}$$

Truncating the SVD (PCA) based on dominant singular values... we get the dominant subspace matrix

$$E_{sub} = U_{truc} \Sigma_{red} V_{truc}^T \tag{5}$$

Now to get the projection, we multiply,

$$P = D_p \cdot E_{sub} \tag{6}$$

It is important to note here that this *dominant subspace* might contain multiple topics too depending on how dispersed the text actually is. We can now extract the column vectors and find their lengths. We will select all sentences

from the original text that has a significant (over half the standard deviation from mean of all) projection. Our work is still proceeding so we are yet to test this on standard data set and check the summarization accuracies. However we have tested it on random Wikipedia samples and some real text samples which were very promising. Table 1 gives performance results on some trials we did with different score functions on texts of increasing length. The experiments were mounted on a nominal dual core laptop with 4 GiBs of RAM. The sentences in each text chunk were cross aligned with each other and the resultant pairwise matrix was used for the summarization process detailed above. Since we could not objectively evaluate the summaries as of now, we simply tabulated some observations from these experimental runs such as the density of the matrix (total percentage of non zero values in the matrix) and the number of sentences in the summary etc. An obvious observation was that the summary remains unaffected with change in the scoring scheme, but the matrix density varies hugely for non relative distances such as the euclidean distance. This can be used empirically to claim that the information capture is coherent through out and not a random pattern. We are able to use such measures (non relative distances) solely due to the fact that all values are normalised as already mentioned. The advantage of using this type of measures is also obvious from their execution time. Their computational complexity (growth rate) is very less compared to the other relative score distances. Since such phenomena is not within our scope of discussion, we will restrict them to these basic observations.

4.1 Spark Implementation

Pairwise analysis is a computationally expensive algorithm as the growth rate complexity is in $O(n^2)$ where n is the number of sentences in the given text. But the algorithm can be easily distributed using a map-reduce programming model that gives high coarse-grained distribution. Text analysis has already been established as a *Big Data* domain; so it is only fitting that we plan ahead and implement this model too on a Big data framework. We have chosen Apache Spark[®] [23] framework for the same. Spark promotes in memory distributed computing that accelerates the execution 20 to 100 times over traditional Hadoop[®] Mapred[®]. Spark interfaces with Java, Python, R and naturally with Scala; our code is fully written in Scala which has a neat functional type syntax and is not verbose like Java. Scala programs run on Java Virtual Machines (JVMs), so they can share all the Java libraries. Ours is mostly a standalone spark code in a Maven project. This can be easily built using Apache Maven[®] and deployed locally or remotely to a decent Spark cluster. We do not use any other library packages except `spark-core` and `spark-mllib`. We employ Spark's distributed data structure called Resilient Distributed Datasets (RDDs) to hold and transform all textual information which will automatically distribute the pairwise computations. Figure 5 shows an entire map of the Spark's computation of the extractive high content summary through various map and filter transformations and reducing to an index of high content sentences.

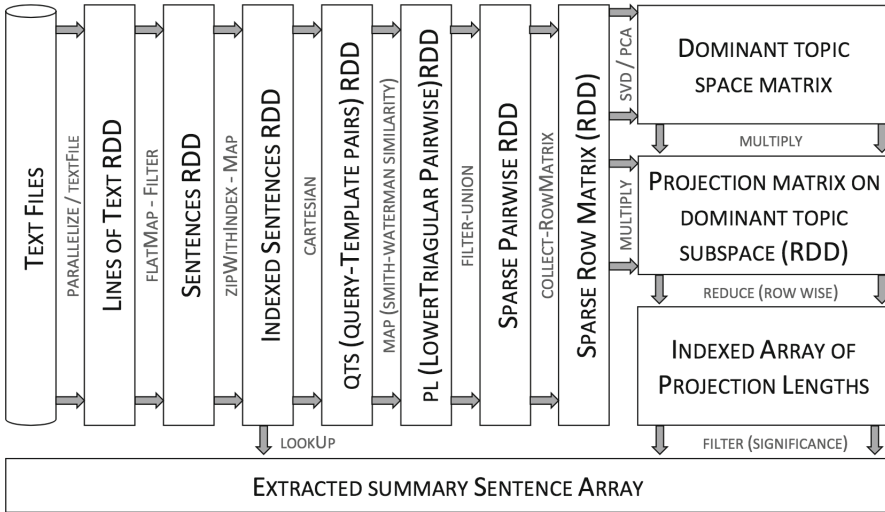


Fig. 5. This is the flow of map-reduce type transformations from raw text files to computing pairwise matrix and find dominant topic subspace, eventually leading to extracting the high content sentences as summary.

5 Future Work and Inferences

We have presented a major crux of this work here. There are a lot more entailing experiments that needs to be done in order to establish this methodology at par with traditional ones used for text analytics tasks. Pairwise analysis using word embeddings can build topic profiles and then use it to test the topic presence in any text or to quantify the degree to which a topic might be present in a given sentence or phrase. Furthermore we can use this comparison technique to test paraphrases and measure their degree of relatedness. This will be the future direction of our work. The pairwise matrix can also be seen as an adjacency matrix and a TextRank[®] type algorithm can rank sentences based on their semantic popularity with in the text. The real TextRank[®] [15] uses LSA as its main comparison technique where we are hopeful that our approach might fetch better results.

References

1. Achananuparp, P., Hu, X., Shen, X.: The evaluation of sentence similarity measures. In: International Conference on Data Warehousing and Knowledge Discovery, pp. 305–316. Springer (2008)
2. Amiri, H., Resnik, P., Boyd-Graber, J., III, H.D.: Learning text pair similarity with context-sensitive autoencoders. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 1882–1892 (2016)
3. Ashwini, B., Menon, V.K., Soman, K.P.: Prediction of Malicious Domains Using Smith Waterman Algorithm, pp. 369–376. Springer, Singapore (2016)

4. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **12**, 2493–2537 (2011)
5. 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*, Association for Computational Linguistics, pp. 340–348 (2010)
6. Gomaa, W.H., Fahmy, A.A.: A survey of text similarity approaches. *Int. J. Comput. Appl.* **68**(13), 13–18 (2013)
7. Gracia, J., Mena, E.: Web-based measure of semantic relatedness. In: *International Conference on Web Information Systems Engineering*, pp. 136–150. Springer (2008)
8. Hassanzadeh, H., Groza, T., Nguyen, A., Hunter, J.: Uqeresearch: semantic textual similarity quantification. In: *SemEval-2015*, p. 123 (2015)
9. He, H., Lin, J.: Pairwise word interaction modeling with deep neural networks for semantic similarity measurement. In: *Proceedings of NAACL-HLT*, pp. 937–948 (2016)
10. He, H., Gimpel, K., Lin, J.J.: Multi-perspective sentence similarity modeling with convolutional neural networks. In: *EMNLP*, pp. 1576–1586 (2015)
11. Irving, R.W.: Plagiarism and collusion detection using the smithwaterman algorithm. Technical report, University of Glasgow, Department of Computer Science (2004)
12. Jensen, A.S., Boss, N.S.: Textual similarity: comparing texts in order to discover how closely they discuss the same topics. B.S. thesis, Technical University of Denmark, DTU, DK-2800 Kgs. Lyngby, Denmark (2008)
13. Kågebäck, M., Mogren, O., Tahmasebi, N., Dubhashi, D.: Extractive summarization using continuous vector space models. In: *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, EACL, Citeseer, pp. 31–39 (2014)
14. van der Maaten, L., Hinton, G.E.: Visualizing high-dimensional data using t-sne. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008)
15. Mihalcea, R., Tarau, P.: Textrank: bringing order into texts. In: Lin, D., Wu, D. (eds.) *Proceedings of EMNLP 2004*, Association for Computational Linguistics, pp. 404–411. Barcelona (2004). <http://www.aclweb.org/anthology/W04-3252>
16. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems, NIPS 2013*, pp. 3111–3119. Curran Associates Inc., USA (2013)
17. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543 (2014)
18. Ramage, D., Rafferty, A.N., Manning, C.D.: Random walks for text semantic similarity. In: *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, Association for Computational Linguistics, pp. 23–31 (2009)
19. Sanborn, A., Skryzalin, J.: Deep learning for semantic similarity. CS224d: Deep Learning for Natural Language Processing Stanford, Stanford University, CA (2015)
20. Smith, T., Waterman, M.: Identification of common molecular subsequences. *J. Mol. Biol.* **147**(1), 195–197 (1981)

21. Turian, J., Ratinov, L., Bengio, Y.: Word representations: a simple and general method for semi-supervised learning. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, ACL 2010, Stroudsburg, PA, USA, pp. 384–394 (2010)
22. Zaharia, M., Chowdhury, M., Franklin, M.J., Shenker, S., Stoica, I.: Spark: cluster computing with working sets. In: Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing, USENIX Association, HotCloud 2010, Berkeley, CA, USA, p. 10 (2010)
23. Zaharia, M., Xin, R.S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M.J., Ghodsi, A., Gonzalez, J., Shenker, S., Stoica, I.: Apache spark: a unified engine for big data processing. *Commun. ACM* **59**(11), 56–65 (2016). doi:[10.1145/2934664](https://doi.org/10.1145/2934664)

Illuminant Color Inconsistency as a Powerful Clue for Detecting Digital Image Forgery: A Survey

Divya S. Vidyadharan^{1,2,3(✉)} and Sabu M. Thampi⁴

¹ College of Engineering-Trivandrum, Trivandrum, Kerala, India
divya.s.vidyadharan@ieee.org

² LBS Centre for Science and Technology, Trivandrum, Kerala, India

³ University of Kerala, Trivandrum, Kerala, India

⁴ Indian Institute of Information Technology and Management-Kerala, Trivandrum, India
sabu.thampi@iiitmk.ac.in

Abstract. Digital images capture our attention and are retained in our memory for longer than other sensory perceptions. Despite numerous instances of image forgery, still, people tend to believe digital images. At the same time, digital investigations reveal an increasing trend of image forgery with illicit purposes. Image editing operations that lead to forgery always leave traces. Investigators rely upon these traces for detecting an image forgery. Researchers are trying to detect image forgery by devising techniques that exploit the traces present in forged images. Recently, illuminant color, the color of the scene illumination present in the image that hints the illumination prevailed at the time of image capture is studied as potential evidence for image forgery. In this survey, we explore the evolution of illuminant color based image forgery detection. This survey provides a brief description of different illuminant color estimation approaches employed in image forgery detection followed by a detailed review of existing illuminant color inconsistency based forgery detection techniques. The major contribution of this survey is the elaborate discussion of future research directions to provide insight to researchers.

Keywords: Illuminant color estimation · Illuminant color inconsistency · Image splicing detection · Image forgery detection · Forgery localization · Image forensics

1 Introduction

Digital images are powerful sensory perceptions considering the natural tendency to believe what we see. Image sharing statistics in the social media shows that Snapchat users share 8796 images per second, whereas Whatsapp users share

8102 images per second and Facebook users share 4501 images every second [26]. This reveals how deeply digital images have ingrained in our daily lives. People are conditioned to believe in images even though an image can be altered effortlessly. Different types of image manipulations include copy-move forgery, image filtering, and image splicing. In copy-move forgery, an image region is copied and pasted onto the same image to enhance the impact of an image. For example, consider an image where a few image regions containing vehicles are copied and pasted on to the same image for depicting heavy traffic. Image filtering is often carried out to enhance the image quality, but sometimes it may alter the original meaning of the image. For example, the color of a vehicle getting altered in a crime scene photograph. In image splicing, an image region from another image is copied and pasted to another image. For example, an image region containing a vehicle being added onto another picture. Examples of image forgery are shown in Fig. 1. Redi et al. have given a detailed discussion regarding the impact and importance of image forgery detection [29].

Just like any act of crime, image alterations also leave some pieces of evidence. These are the traces left by the image processing operations carried out during alteration. Even saving a JPEG image again in JPEG format after altering the image contents can introduce JPEG double compression artifacts [37]. Forensic analysis of digital images involves the detection and analysis of the evidence left behind during a forgery. Various pieces of evidence considered include similarities in pixel-wise regularities (for detecting copy-paste forgery) and interpolation pattern inconsistencies (for detecting image splicing). Image splicing introduces several pieces of evidence including inconsistencies in the underlying camera noise pattern and inconsistencies in scene illuminant color.

Several surveys have been published in the broad area of digital image forensics [3, 17, 28, 33]. In this paper, we survey the existing image splicing detection techniques that exploit inconsistencies in illuminant color. This work elaborates the different illuminant representation models and clarifies how illuminant color is computed in each model. We hope that the discussion on the background would help readers gain a better understanding of the underlying process of illuminant color estimation in each technique.

1.1 Motivation

Recently, researchers have started analyzing illumination in a scene during image capture as evidence to expose an image forgery. Compared to the independent research advancing in the directions of illuminant color estimation, and image forgery detection separately, the illuminant color estimation based image forgery detection is in the evolving stage. So far, research bottlenecks and opportunities in illuminant color estimation based forgery detection have been disclosed sparingly in research theses and specific papers only. This motivated us to carry out a survey of existing work based on illuminant color inconsistency and to attempt giving hints to researchers regarding possible future research directions.

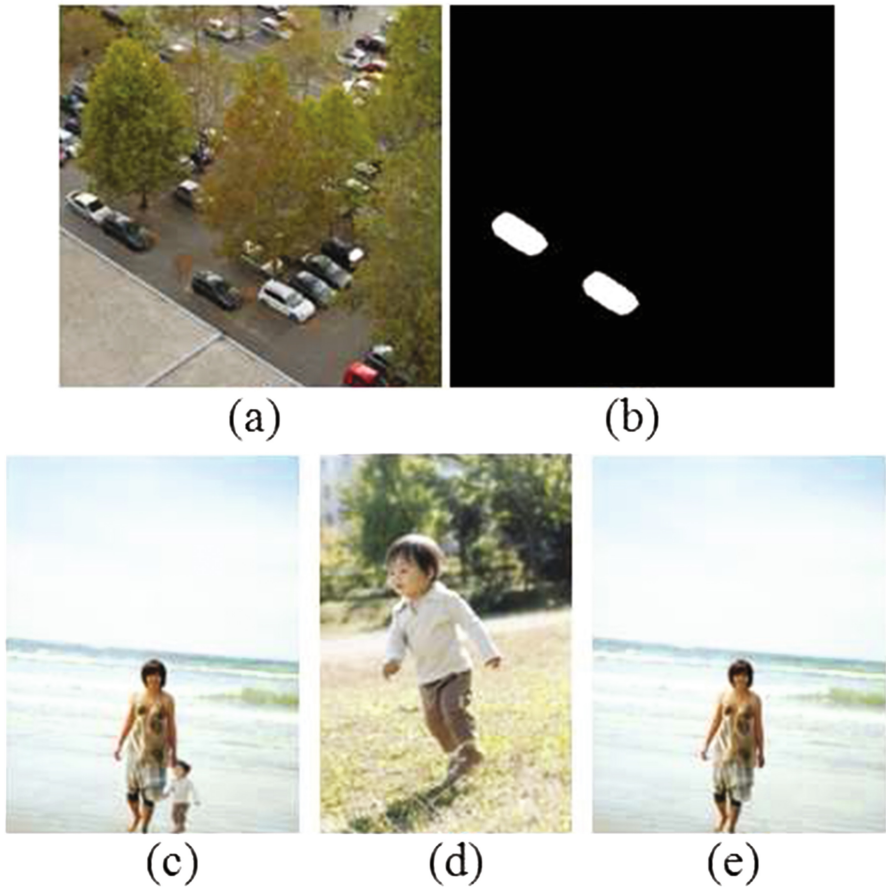


Fig. 1. (a) An example of copy-move forgery. (b) The mask showing copy moved regions. Both the image and the mask are taken from CoMoFoD dataset [39]. (c) An example of image splicing. (d) and (e) Source image and Destination image for creating the spiced image in (c). Images shown in (c) and (e) are taken from CASIA V2.0 dataset [14].

1.2 Contribution

The main contributions of this survey are

- A discussion on illuminant color estimation approaches for a better understanding of illuminant color inconsistency based techniques.
- A survey of existing illuminant color inconsistency based image forgery detection techniques.
- A detailed listing of future research directions including the need for specific datasets and the possibility of applying new technologies.

The rest of the paper is organized as follows. Section 2 introduces the basic concept of scene illumination. Section 3 explains different illumination estimation approaches employed in existing illuminant color inconsistency based forgery detection techniques. Section 4 examines various illuminant color based forgery detection techniques. Finally, Sect. 5 elaborates future research directions followed by a conclusion in Sect. 6.

2 Scene Illumination

Scene illumination represents the illumination prevailed in the scene at the time of image capture. The scene illumination in an outdoor scene will be uniform whereas the scene illumination in an indoor scene will be non-uniform as indoors are often lit up by a mix of multiple light sources. The scene illumination influences the color or pixel value recorded by the camera sensor. Hence, in an image, the perceived color is not the actual color of the object, instead, a combination of the object color and the color of scene illumination. The color of scene illumination is termed as the illuminant color.

Humans have the ability to see objects in their actual color irrespective of illuminant color. This capability of the human visual system is known as color constancy. Incorporating color constancy for computer vision applications is an active research area. For object recognition purposes, the illuminant color is estimated and later removed to get the actual color of the object. When an image is altered by copy-pasting a region from another image, there will be a mismatch in the illumination of the copy-pasted region with the rest of the image. Therefore, the inconsistency in illuminant color across an image can be considered as a clue for detecting an image forgery. If an illuminant color estimated from a suspect region is different from the illuminant color estimated from the rest of the image, possibly the suspect region could have been copy-pasted from another image captured with a different scene illumination. Illuminant color is usually estimated by following any of the illuminant color estimation approaches discussed in Sect. 3.

3 Different Approaches for Estimating Scene Illuminant Color

Several illuminant color estimation techniques are available. For further reading, please refer to Gijssen et al.'s survey [21] where authors have surveyed and evaluated various color constancy techniques. In our survey, the discussion is restricted to statistics based and physics based approaches as these two approaches are the common illuminant color estimation approaches employed in current image forgery detection techniques.

3.1 Statistics-Based Approach

In statistics-based approach, the techniques rely upon the color distribution present in the image and are influenced by the number of colors present in the image. For example, the traditional Gray-world [5] assumes that the average color in an image is gray. Hence, any deviation from this gray is contributed by the illuminant color. Another generalized model is Generalized Gray-Edge (GGE) assumption proposed by Van De Weijer et al. [40]. The GGE assumes that the average color of edges in an image is gray. In this model, the illuminant color is computed by taking the integral of derivatives of pixels in an image.

3.2 Physics-Based Approach

Physics-based techniques are based on the understanding of physical properties of light reflection and hence perform well even if the number of colors in an image are few [18]. A popular Physics-based illuminant color estimation approach is based on the Dichromatic Reflection Model (DRM) [34]. According to DRM, homogeneous objects (objects with a uniform surface) show only the interface reflection and inhomogeneous objects show both the interface and the body reflection [38].

In DRM, the light reflected from an inhomogeneous dielectric surface is considered as a combination of specular reflectance (specular highlights, for e.g., the bright cheek region in a facial image) and diffuse/body reflectance (light reflected from the surface albedo/matte). Specular reflectance is also known as interface reflectance since it is the part of the light immediately reflected from the surface causing specular highlights. The estimation of illuminant color in DRM is explained here.

Tan et al. have defined chromaticity (normalized RGB) as the ratio of an RGB component to the sum of R, G, and B components [38]. When pixels from a uniformly colored object are plotted in a chromaticity - intensity space, the interface (specular) pixels will appear as a varying cluster, whereas the body (diffuse) pixels appear as a straight vertical line showing that the diffuse pixels are independent of the image intensity, as shown in Fig. 2 [38].

The varying cluster of specular pixels can be clearly understood in the Inverse Intensity-Chromaticity (IIC) space where the x-axis represents Inverse Intensity, defined as,

$$\text{Inverse Intensity} = 1 / \sum I_i(x) \quad (1)$$

and the y-axis represents chromaticity. When pixels from a uniformly colored surface are projected onto the IIC space, we get the illuminant chromaticity as illustrated in Fig. 3 [38].

To compute the illuminant chromaticity, the IIC space is transformed into Hough space. In the Hough space, the x-axis represents the illuminant chromaticity and y-axis represents the image chromaticity as shown in Fig. 4(a). The illuminant chromaticity with the maximum number of line intersections is

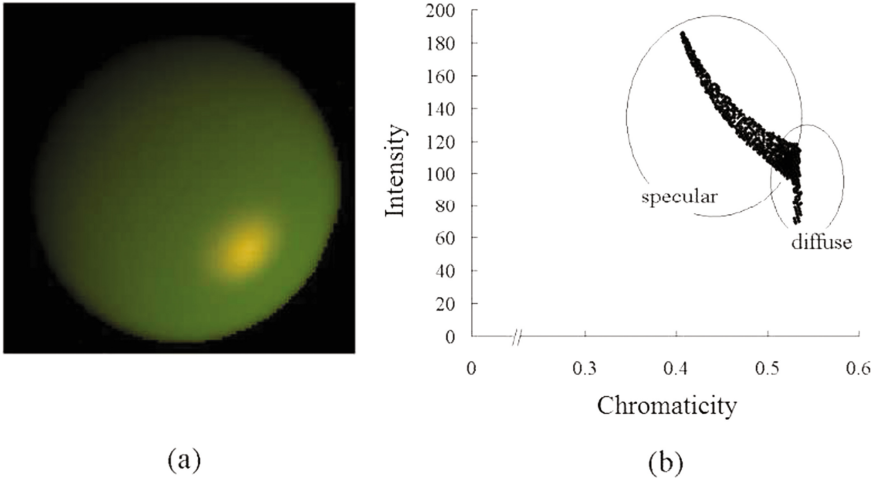


Fig. 2. (a) A green colored synthetic object. (b) The projection of specular and diffuse pixels (green plane) in the chromaticity-intensity space (right) [38] (Reprinted from Tan, R.T., Nishino, K., Ikeuchi, K.: Color constancy through inverse-intensity chromaticity space. *Journal of Optical Society of America A* 21(3), 321–34 (2004)).

taken as the illuminant chromaticity, as illustrated in Fig. 4(b). To get the illuminant color in RGB color space, the illuminant chromaticity computation is carried out in R, G, B channels separately.

Ideally, in an authentic image, the illuminant chromaticity will be consistent throughout the image pixels. But, during image splicing, where an image region

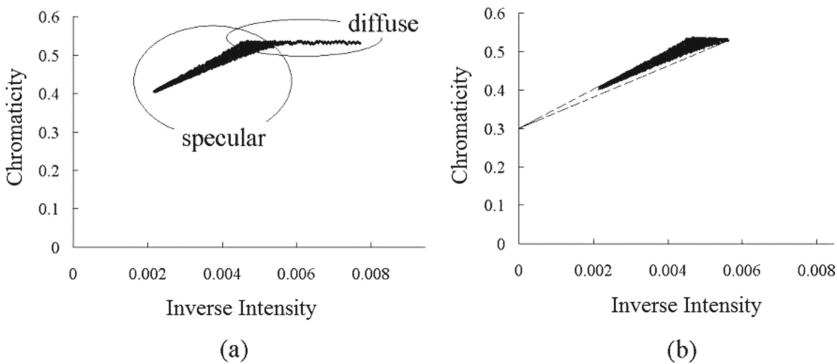


Fig. 3. (a) Inverse Intensity-Chromaticity space showing specular and diffuse pixel clusters. (b) The specular cluster extension pointing towards the illumination-chromaticity (green plane) in the y-axis [38] (Reprinted from Tan, R.T., Nishino, K., Ikeuchi, K.: Color constancy through inverse-intensity chromaticity space. *Journal of Optical Society of America A* 21(3), 321–34 (2004)).

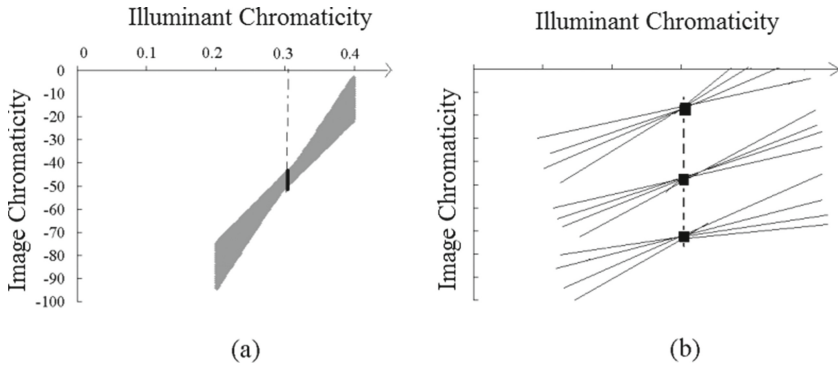


Fig. 4. (a) Hough space. (b) The intersection of lines in Hough space [38] (Reprinted from Tan, R.T., Nishino, K., Ikeuchi, K.: Color constancy through inverse-intensity chromaticity space. *Journal of Optical Society of America A* 21(3), 321–34 (2004)).

is copy-pasted from another image with a different illuminant chromaticity, the illuminant distribution across the spliced image will become inconsistent.

4 Illuminant Color Inconsistency Based Image Forgery Detection Techniques

Recently, researchers have started analyzing illuminant color inconsistency for image forgery detection. Illuminant color based forensics is challenging because (i) most of the existing illuminant color estimation methods assume single illumination whereas the real world scenes can be multi-illuminated [4], (ii) illuminant color estimation is an ill-posed problem, and (iii) illuminant color comparisons can only be carried out on similar materials since the material properties affect the surface reflection.

In general, the illuminant color comparison is restricted to similar material surfaces. In this survey, illuminant color inconsistency based techniques are grouped into two, since research is happening in parallel in two directions based on objects analyzed. In the first direction, the forgery detection is carried out by analyzing the illuminant color from similar objects [6, 16, 20, 43]. In the second direction, techniques concentrate on detecting forgery by analyzing skin regions [13, 19, 31, 41]. This classification has significance in the digital forensics domain since a lot of image forgery cases are being reported where human skin regions are copy-pasted.

4.1 Forgery Detection Techniques Analyzing Object Regions

Gholap and Bora devised an illuminant color estimation based forgery detection based on the dichromatic reflection model [20]. All the R, G, B values of pixels

in the specular highlight regions are arranged as a matrix and Principal Component Analysis (PCA) is carried out by Singular Value Decomposition (SVD). The eigenvectors for the two significant eigenvalues from the two principal components constitute the dichromatic plane. The dichromatic planes are projected as lines in normalized r-g chromaticity space and the point of the intersection of the colors indicates illuminant color. In an authentic image, the different dichromatic lines estimated from different objects intersect at the same point.

But, if the image is forged by copy-pasting different regions from different images, then the dichromatic lines obtained may not intersect at the same point indicating forgery. This method assumes that there is only a single light source in the image.

Cao et al. developed an image splicing detection technique based on the differences in the local color statistics, and the difference in illuminant color between the suspect region and the background region [6]. The method is based on the fact that local color statistics will be consistent in natural authentic images. In this method, the image is segmented into a foreground region and a background region, and color histograms are extracted from the background region and the foreground object regions separately. The distance between the foreground and background histogram is computed using histogram distance measures such as Chi-square distance and Kullback-Leibler (K-L) distance, constituting the first set of features. The second set of features is obtained from illuminant color inconsistency between the foreground and background regions.

To obtain the difference in illuminant color, a new method of illuminant color estimation is proposed. Here, the illuminant color is estimated as the average color of near-white pixels. Near-white pixels are pixels that exhibit near-zero color difference and near-white luminance. The average color differences in these near-white pixels between the foreground and background regions are computed in both U and V planes in YUV color space. Finally, the features are fed to an SVM classifier with RBF kernel. Test dataset contained 180 realistic, 360 unrealistic forged images and 540 real images. Different color spaces such as RGB, YUV, HSV, XYZ, $L^*a^*b^*$, and $L\alpha\beta$ are used. Experiments are conducted with different color spaces and different histogram distance measures. The K-L distance measure in the HSV color model gave optimum results. The results show that this illuminant color-based method obtained an Average Precision of 56% at a low False Positive (FP) rate of 0.2%.

Wu and Fang proposed another illuminant color inconsistency based image splicing detection technique assuming single illumination [43]. Here, the image is divided into overlapping blocks. The method makes use of three illuminant color estimating techniques such as the Grey-Shadow, the first-order Grey-Edge and the second order Grey-Edge algorithm, represented by Generalized Grey Edge framework proposed by Van De Weijer et al. [40]. The most suitable algorithm for each block is adaptively selected using a maximum likelihood classifier proposed by Gijsenij et al. [22] based on block properties such as color distribution and color edges. The illuminant color for each block is estimated using the selected algorithm. The estimated illuminant color is compared with reference blocks. The

comparison is carried out by computing the angular error between the illuminant color in each block and the reference block. If this angular error is greater than a threshold, the corresponding block is considered as spliced.

The image database by Ciurea and Funt is used for classifier training and for determining the block size [10]. A block size of 30×30 is selected, since increasing the block size further reduced localization accuracy. The adaptive illuminant algorithm selection achieves an accuracy of 94% when tested on the database. The angular error threshold is determined using 100 spliced images taken from the CASIA image tampering database [14]. A detection accuracy of 75% is obtained when the angular error threshold is set to 7.

Fan et al. overcame the need for manual selection of reference objects in their work [16]. Here, the image is divided into vertical and horizontal bands. Illuminant colors are estimated using five algorithms such as Grey-World, Max-RGB, Shades of Grey, first-order Grey-Edge and Second order Grey-Edge. Thus, five illuminant color values are obtained for each band. For each illuminant estimation algorithm, two reference illuminant colors are obtained by taking the median of vertical and horizontal bands separately. If the distance between the illuminant color of a band and the reference illuminant is greater than a preset value, that band is considered as spliced. For each illuminant color estimating algorithm, the intersection of bands marked as spliced is represented as a detection map. Finally, the spliced region is detected by the intersection of all bands previously considered as spliced.

Experiments are conducted on two sets of images taken from CASIA V2.0 [14]. The first dataset contains images without reference objects. Based on the image contents, this dataset is divided into four categories such as ‘People’, ‘Animals’, ‘Plants’, and ‘Objects’. The second data set contains reference objects, where manual marking is required to identify three objects including one object belonging to the spliced region in the image. The proposed method obtained highest True Positive Rate (TPR) of 90.00% in the Plants group and the lowest False Positive Rate (FPR) of 20.00% in the Objects group and highest Accuracy (ACC) of 76.62% in the Objects group. Overall TPR, FPR and ACC are 50.75%, 26.34% and 67.59%. The performance of the method in the second dataset with the reference object is compared with two other illumination based splicing detection methods such as Method based on Dichromatic Line (MDL) [24] and Method based on Illumination Map (MIM) [31]. Experimental results show that though MIM gave the highest TPR of 60.00%, the proposed method gave lowest FPR of 19.58% and the highest ACC of 76.69%. When the computation time is compared, Fan et al.’s method required 713.66 s whereas MDL and MIM required 296.66 s and 640.14 s respectively.

Illuminant inconsistency based techniques address two kinds of issues, such as forgery detection and forgery localization. In forgery detection, the technique classifies an image as either a forged or an authentic image whereas in forgery localization, the region of forgery is identified. Another point considered while comparing the techniques are the underlying assumption regarding illumination. Certain methods discussed are based on the uniform single illuminant source

Table 1. Summary of forgery detection techniques analyzing object regions.

Methods	Classification or localization	Illumination model	Reference or no-reference	Single or multi-illuminant
Gholap and Bora [20]	Classification	Dichromatic reflectance model (DRM)	No reference region needed	Single illuminant
Cao et al. [6]	Spliced region localization	-	Reference foreground region needed	-
Wu and Fang [43]	Spliced region localization	Statistical	Reference needed	Single Illuminant
Fan et al. [16]	Spliced region localization	Statistical	No reference region needed	Can also handle multi-illuminant

Table 2. Comparison of performance forgery detection techniques analyzing object regions.

Method	Dataset	Performance metric	Results	
Cao et al. [6]	Author’s dataset	Precision (Average)	56%	
		FPR	0.2%	
Wu and Fang [43]	100 images in CASIA V1.0	Detection accuracy	75%	
Fan et al. [16]	Two groups from CASIA V2.0	First group (4 sub-categories)	TPR	50.75%
			FPR	26.34%
			Accuracy	67.59 %
		Second group	TPR	52.31 %
			FPR	19.58 %
			Accuracy	76.69 %

assumption, whereas some other methods are based on the multi-illuminant assumption. Among the four techniques described above, the technique proposed by Gholap and Bora [20] classifies an image as spliced or authentic, whereas techniques by Wu and Fang [43], Cao et al. [6] and Fan et al. [16] perform forgery localization. For the techniques proposed by Cao et al. [6] and Wu and Fang [43], a reference region is to be specified to detect the inconsistency in illuminant color, whereas techniques proposed by Gholap and Bora [20], and Fan et al. [16] do not require any reference region.

Table 1 summarizes the methods discussed in Sect. 4.1. A comparison of the performance of the methods proposed by Cao et al. [6], Wu and Fang [43], and Fan et al. [16] is given in Table 2. All the techniques for which the experiments are conducted on a dataset are included in Table 2.

4.2 Forgery Detection Techniques Analyzing Facial Skin Regions

While altering an image, a human facial region may be copied from an image taken in a different lighting environment. This introduces a discrepancy at the copy-pasted facial region compared to the authentic facial regions. Among the various forgery detection techniques that considered facial skin regions [8, 13, 19, 25, 31, 41, 42], the techniques that considered the illuminant color inconsistency are discussed here.

Riess and Angelopoulou proposed a method based on the dichromatic reflection model that identifies the variation in illuminant color using two maps generated from the image - an illuminant map and a distance map [31]. Here, the image is segmented into sub-regions based on color similarity. Each sub-region is again partitioned into small patches and an illuminant color is estimated from each small patch. Illuminant color is computed using the Inverse Intensity Chromaticity (IIC) space described in Sect. 3.2. In the IIC space, the diffuse pixels in the small sub-region will form a horizontal line, whereas the bright specular pixels point toward the illuminant color in the Chromaticity axis. Pixel groups that satisfy the two constraints in the IIC space, such as a constraint on the shape of pixel distribution, and another constraint on the slope of pixel distribution, are only considered. Illuminants are estimated from these small pixel patches, and finally, an illuminant is selected through majority voting. The illuminant color thus obtained is used to generate an illuminant map where each sub-region is colored with selected illuminant color. The distance map is generated by representing the deviation of illuminant color computed from specially selected sub-regions to the rest of the sub-regions.

Both the illuminant map and the distance map show the inconsistency in illuminant color in the altered region. An example is shown in Fig. 5. A manual



Fig. 5. (a) Forged image with the third person copy-pasted. (b) Illuminant map. (c) Distance map. Illuminant map and distance map clearly shows the third face as an inconsistent region [31] (Reprinted from Riess, C., Angelopoulou, E.: Scene illumination as an indicator of image manipulation. In: Information Hiding. vol. 6387, pp. 66–80 (2010) with permission from Springer).

examination of the illuminant map and distance map reveals the copy-pasted image region. The advantage of this method is that it calculates the illuminant color at a local region and hence works well on real-world multi-illuminant images.

Carvalho et al. have proposed a machine learning based method [13] that automates the previous image splicing detection method proposed by Riess and Angelopoulou [31]. This method analyzes facial skin pixels for detecting an image forgery. The method consists of five stages.

In the first stage, two variants of illuminant maps are generated after partitioning the image into pixels of similar color. One variant, the IIC based illuminant map is generated as proposed by Riess and Angelopoulou [31]. The second variant is the statistics-based Generalized Gray World (GGW) illuminant map. For generating GGW illuminant map, the illuminant color for each small pixel group is estimated using the method proposed by Van De Weijer et al. [40]. In the second stage, the facial regions from the illuminant maps are extracted by the user specifying a bounding box around the face. In the third stage, a feature set consisting of texture and edge descriptors are extracted. Edge features are generated using a new edge-based Histogram of Gradient (HOG) descriptor based on the HOG-descriptor [12]. For texture features, Statistical Analysis of Structural Information (SASI) descriptor is used [7]. Both edge and texture features are extracted from the IIC and GGW illuminant maps. The method identifies an image as tampered if any of the face pairs in the image is inconsistently illuminated. Thus, in the fourth stage, all face pairs are considered and the features from a face pair are concatenated. In the final fifth stage, an image is categorized as tampered if any of the two faces are identified as inconsistent. The SASI-Gray-World, SASI-IIC, HOGedge-IIC and HOGedge-GGW features are fed to a Support Vector Machine (SVM) classifier independently. Then, the SVM meta-fusion combines the output of all the independent classifiers as a combined feature set. This new feature set is fed to another SVM classifier to categorize the image as tampered or original.

In this work, Carvalho et al. introduced two datasets DSO-I and DSI-1. DSO-I contains 200 images (100 original and 100 spliced) with a resolution of $2,048 \times 1536$ pixels. The DSI-1 dataset consists of 25 authentic and 25 tampered images downloaded from the internet. When the meta-fusion SVM classifier is tested on DSO-I dataset, it obtained an overall Area Under the Curve (AUC) of 86.3% whereas a manual evaluation of the same dataset achieved only 38.3% on tampered images. The DSI-1 dataset is used for a cross-database experiment on the classifier trained with DSO-I and obtained an AUC of 82.6% indicating generalization to images from other sources as well.

Francis et al. devised illumination based forgery detection from human skin highlight pixels [19]. The proposed method works as follows. The input image is segmented into facial regions of different persons present in the image. For each person, pixels in nose tip are selected (can be done manually or automatically using any face detection technique). Principal Component Analysis (PCA) is performed on the sorted pixels starting from the darkest pixel for estimating

the body reflection vector. The PCA is performed on the sorted pixels starting from the brightest pixel to obtain specular reflection vector. The direction of specular reflection vector is mapped onto the RGB chromaticity space. This direction gives the estimate of the illuminant color. In the normalized RG space, the chromaticity coordinates of illuminant colors obtained for different persons are plotted. The Euclidean distance between points is calculated, and if the distance measure is greater than a threshold then it indicates forgery. This method requires frontal facial regions for estimating the illuminant color from nose tip highlights.

Carvalho et al. extended their previous work [13], by considering more color spaces in addition to YCbCr, and by using a more powerful classifier fusion and selection method [8,9]. In this method, both GGW and IIC based illuminant maps are generated from the color segmented input image. The facial regions are represented in four different color spaces such as HSV, Lab, YCbCr, and RGB since different features are highlighted in different color spaces. Various visual properties of the image such as texture, shape, and color are extracted and represented as image descriptors. A combined image descriptor representing the illuminant map, color space, and visual properties are computed. A feature vector is then obtained for a pair of faces by concatenating the image descriptor for a pair of faces. An optimum combination of these feature vectors is then selected and classified through a classifier and fusion technique. A classification rate of 94% with a reduction of 72% of error from the previous method [13] is achieved.

In addition to forgery detection, forgery localization is also performed by computing the probability of a face to be spliced using an SVM classifier after an image is classified as spliced. Forgery localization is based on the finding that the difference in illuminant maps (GGW and IIC) for a spliced facial region is higher than that for an original face. For forgery localization, an SVM classifier with various color descriptors such as color correlograms [23], Border/ Interior pixel Classification (BIC) [35], color coherence vectors [27] and local color histograms [36] are used, and obtained a detection accuracy of 76%, 85%, 83% and 69% respectively.

Vidyadharan and Thampi proposed another forgery localization technique [41] using illuminant maps introduced by Riess and Angelopoulou [31]. Both Generalized Gray-World (GGW) and Inverse Intensity Chromaticity (IIC) based illuminant maps are used. In this technique, the facial regions are extracted manually from illuminant maps. The extracted facial regions in RGB space are converted to gray scale. All the facial regions are arranged as an $M \times N$ matrix where each row represents a facial region as an N dimensional vector. N is the total number of pixels and M is the number of faces in the image. PCA is carried out in this $M \times N$ matrix by decomposing the matrix using Singular Value Decomposition (SVD). The matrices representing facial regions when projected on to the principal component space shows the illumination variance between faces. Facial regions showing similar illumination properties will be grouped together in the principal component axes and the face with dissimilar illumination

properties will be projected as an outlier. When the proposed method is evaluated in images containing three or more faces from DSO-I dataset, the detection accuracy obtained on GGW and IIC illuminant maps are 64% and 62% respectively. For images with three or more faces in the DSI-1 dataset, a detection accuracy of 42% is obtained for both IIC and GGW maps.

Mazumdar and Bora devised an illumination-signature capable of detecting forged images [25]. The illumination-signature, the Dichromatic Plane Histogram (DPH) is based on DRM. From each face, a DPH is generated using 2-D Hough Transform. DPHs obtained from two faces are compared using correlation measure. If the correlation value is higher than a pre-specified threshold, the illumination is considered as consistent between the faces and hence the image is identified as an authentic image. On the other hand, if the correlation value is below the threshold, the illumination is considered inconsistent and hence the image is identified as forged. The method is evaluated on a combination of subset of images from DSO-I and DSI-1 datasets, and a new dataset created by the authors - Face Splicing Detection (FSD) dataset. The proposed method obtained an AUC of 91.2% when tested on the combined dataset containing 55 spliced and 55 authentic images. Further experiments conducted with images compressed in JPEG quality factors, 90, 80, and 70 gave AUC values of 90.8%, 90.6% and 89.6% respectively. This shows that the proposed method is robust to JPEG compression.

Table 3. Summary of forgery detection techniques analyzing facial skin regions.

Methods	Classification or localization	Model	Approach	Single or multi-illuminant	Remarks
Riess and Angelopoulou [31]	Localization	DRM	Non-machine learning	Multi-illuminant	Pioneering work on illuminant maps no training needed
Carvalho et al. [13]	Detection	Both statistical and DRM	Machine learning - SVM metafusion	Multi-illuminant	First automated technique introduced DSO-I and DSI-1 datasets
Carvalho et al. [8]	Detection and localization	Both statistical and DRM	Machine learning - multi classifier kNN	Multi-illuminant	State-of-art
Francis et al. [19]	Localization	DRM	Non-machine learning	-	No training needed
Vidyadharan and Thampi [41]	Localization	Both statistical and DRM	Non-machine learning	Multi-illuminant	No training needed
Mazumdar and Bora [25]	Localization	DRM	Non-machine learning	Multi-illuminant	No training needed

The illuminant color inconsistency based forgery detection methods that considered facial regions are studied based on the task - Forgery Detection/Localization, the approach followed - machine learning/non-machine learning, the assumption regarding the illuminant - single/multi-illuminant. A summary of the

Table 4. Comparison of forgery detection techniques analyzing facial skin regions.

Method	Dataset	Approach	Task	Performance metric	Results
Carvalho et al. [13]	DSO-I	Machine learning	Detection	Area under the curve (AUC)	86.3%
Carvalho et al. [8]	DSO-I	Machine learning	Detection	Detection accuracy	94%
			Forgery localization	Detection accuracy	85%
Vidyadharan and Thampi [41]	A sub set of images from DSO-I	Non-machine learning	Forgery localization	Detection accuracy	62% (on IIC, on a subset)
					64% (on GGW, on a subset)
Mazumdar and Bora [25]	A combination of sub set of images from DSO-I and DSI-1	Non-machine learning	Detection	AUC	91.2% (on a subset)

techniques discussed in Sect. 4.2 is given in Table 3. A performance comparison of the methods discussed in Sect. 4.2 is given in Table 4. Only techniques with experimental results available on a dataset are included in Table 4.

5 Future Research Directions

Illuminant inconsistency based forgery detection is an evolving research area considering the lack of multi-illuminant estimation techniques and proper dataset. Here, we discuss the future research opportunities exposed by other researchers along with the directions revealed during our literature survey.

The effect of the camera's inbuilt color constancy algorithm. Riess has observed that it is not possible to tell how the illuminant color is being affected by the camera's own color constancy methods [30]. Riess clearly mentioned the need for a proper dataset that includes images captured using different camera models with a color chart present in each image to carry out illuminant color analysis. According to Riess, this kind of dataset could help in exploring interpolation patterns or camera response function for forensic analysis of digital images.

The effect of JPEG compression, noise, and blur. Another research direction is the study of the effect of compression schemes such as JPEG in the illuminant color based techniques, as mentioned in the work of Carvalho et al. [13]. Current state-of-art methods work well on uncompressed data compared to JPEG compressed images. Hence, how the JPEG compression process and how the presence of compression artifacts affect the illuminant maps need to be explored. Similarly, the effects of noise, camera out-of-focus, and blur also require further exploration.

The effect of known illuminant-color-variation and skin-tone variation. Carvalho et al. in the recent work [8] noticed that, in the future, three kinds of experiments can be considered. First, images captured at known lighting can

be used and the proposed detector can be tested with image pairs that vary in illumination by a known amount. Secondly, experiments that analyze the distribution of illumination in different people in an image are to be carried out. Finally, the influence of different skin tones can be studied.

The effect of Fresnel rectification of skin pixels. The illuminant color estimation using the Inverse-Intensity Chromaticity space [38], assumes that the color of specular pixels is the color of the illuminant. This assumption is known as Neutral Interface Reflection (NIR). However, the geometry of the scene and the refractive index of the surface affect the specular pixels. This wavelength dependent refractive index, the color of the object, and the geometry are captured as a function of wavelength known as Fresnel term in the reflectance model proposed by Cook and Torrance [11]. The Fresnel effect is neglected in the NIR based model. Eibenberger and Angelopoulou found out that the Fresnel effect, when ignored introduces an error in specular based illuminant color estimation methods [15]. Eibenberger and Angelopoulou showed that rectification for this illuminant color shift in human skin pixels can improve the illuminant color estimation by 30%. In the future, illumination estimation from human skin regions should consider this correction as well.

Application of recent illuminant color estimation methods. Although, Riess and Angelopoulou's pioneering work on illumination representation [31] and subsequent works by Carvalho et al. [8,13] take care of multi-illumination, researchers can as well consider recent multi-illuminant estimation methods such as the method proposed by Beigpour et al. [1]. Similarly, the method of adaptive color constancy from skin pixels proposed by Bianco and Schettini [2] can be considered for skin-pixel based forgery detection [13]. Also, illuminant estimation from multiple dissimilar surface materials can also be attempted, as in the recent work that overtook the need for similar surface materials for detecting the direction of the light source [32].

Application of deep neural networks. The machine learning based illumination inconsistency detection techniques proposed by Carvalho et al. rely upon texture, edge and color features extracted from images [8,13]. Recently, the feature extraction based computer vision applications are addressed by deep learning techniques. Deep learning techniques can be explored to tackle image forgery detection. Currently, lack of large datasets setback research in this direction.

6 Conclusion

Inconsistency in illumination can be considered as a potential clue while authenticating digital images during a digital crime investigation. This survey gives an overview of illuminant color inconsistency based image forgery detection mechanisms devised recently. The underlying illumination models are explained to help researchers understand the techniques clearly. Illuminant color inconsistency based forgery detection schemes are grouped into two categories based on the type of image regions considered. Since the images with human skin regions are important in many forensic investigations, techniques that deal with

human facial regions are categorized separately. In a nutshell, researchers need to consider the creation of new dataset along with ground truth illuminant color information for each image, explore new research directions that take care of the effect of JPEG compression, noise, blur, and Fresnel effect rectification for skin pixels, and try to incorporate new multi-illuminant estimation techniques.

Acknowledgments. The authors would like to thank the Higher Education Department, Government of Kerala for funding this research and the Department of Computer Science and Engineering, College of Engineering, Trivandrum for providing the facilities.

References

1. Beigpour, S., Riess, C., van de Weijer, J., Angelopoulou, E.: Multi-illuminant estimation with conditional random fields. *IEEE Trans. Image Process.* **23**(1), 83–96 (2014)
2. Bianco, S., Schettini, R.: Adaptive color constancy using faces. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**(8), 1505–18 (2014)
3. Birajdar, G.K., Mankar, V.H.: Digital image forgery detection using passive techniques: a survey. *Digital Invest.* **10**(3), 226–245 (2013)
4. Bleier, M., Riess, C., Beigpour, S., Eibenberger, E., Angelopoulou, E., Tröger, T., Kaup, A.: Color constancy and non-uniform illumination: can existing algorithms work? In: *Proceedings of IEEE Color and Photometry in Computer Vision Workshop*, pp. 774–81 (2011)
5. Buchsbaum, G.: A spatial processor model for object colour perception. *J. Franklin Inst.* **310**(1), 1–26 (1980)
6. Cao, G., Zhao, Y., Ni, R.: Image composition detection using object-based color consistency. In: *2008 9th International Conference on Signal Processing.*, pp. 1186–1189, October 2008
7. Çarkacıoğlu, A., Yarman-Vural, F.: Sasi: a generic texture descriptor for image retrieval. *Pattern Recognit.* **36**(11), 2615–33 (2003)
8. Carvalho, T., Faria, F.A., Pedrini, H., Torres, R.D.S., Rocha, A.: Illuminantbased transformed spaces for image forensics. *IEEE Trans. Inf. Forensics Secur.* **11**(4), 720–33 (2016)
9. Carvalho, T., Pedrini, H., Rocha, A.: Illumination inconsistency sleuthing for exposing fauxtography and uncovering composition telltales in digital images. In: *Workshop of Theses and Dissertations-XXVII SIBGRAPI Conference on Graphics, Patterns and Images*, Rio de Janeiro, RJ, Brazil (2014)
10. Ciurea, F., Funt, B.: A large image database for color constancy research. In: *Color and Imaging Conference*, Society for Imaging Science and Technology, vol. 2003, pp. 160–164 (2003)
11. Cook, R.L., Torrance, K.E.: A reflectance model for computer graphics. *ACM Trans. Graph. (TOG)* **1**(1), 7–24 (1982)
12. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, vol. 1, pp. 886–893, June 2005
13. De Carvalho, T.J., Riess, C., Angelopoulou, E., Pedrini, H., de Rezende Rocha, A.: Exposing digital image forgeries by illumination color classification. *IEEE Trans. Inf. Forensics Secur.* **8**(7), 1182–94 (2013)

14. Dong, J., Wang, W., Tan, T.: Casia image tampering detection evaluation database. In: 2013 IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP), pp. 422–426. IEEE (2013)
15. Eibenberger, E., Angelopoulou, E.: Beyond the neutral interface reflection assumption in illuminant color estimation. In: Proceedings of IEEE International Conference Image Processing (ICIP), pp. 4689–4692 (2010)
16. Fan, Y., Carré, P., Fernandez-Maloigne, C.: Image splicing detection with local illumination estimation. In: 2015 IEEE International Conference on Image Processing (ICIP), pp. 2940–2944, September 2015
17. Farid, H.: Image forgery detection. *IEEE Signal Process. Mag.* **26**(2), 16–25 (2009)
18. Finlayson, G.D., Schaefer, G.: Solving for colour constancy using a constrained dichromatic reflection model. *Int. J. Comput. Vis.* **42**(3), 127–144 (2001)
19. Francis, K., Gholap, S., Bora, P.K.: Illuminant colour based image forensics using mismatch in human skin highlights. In: 2014 Twentieth National Conference on Communications (NCC), pp. 1–6. IEEE (2014)
20. Gholap, S., Bora, P.K.: Illuminant colour based image forensics. In: TENCON 2008–2008 IEEE Region 10 Conference, pp. 1–5, November 2008
21. Gijssenij, A., Gevers, T., van de Weijer, J.: Computational color constancy: survey and experiments. *IEEE Trans. Image Process.* **20**(9), 2475–89 (2011)
22. Gijssenij, A., Gevers, T.: Color constancy using natural image statistics and scene semantics. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**(4), 687–98 (2011)
23. Huang, J., Kumar, S.R., Mitra, M., Zhu, W.J., Zabih, R.: Image indexing using color correlograms. In: Proceedings of 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 762–768, June 1997
24. Mazin, B., Delon, J., Gousseau, Y.: Estimation of illuminants from projections on the planckian locus. *IEEE Trans. Image Process.* **24**(6), 1944–55 (2015)
25. Mazumdar, A., Bora, P.K.: Exposing splicing forgeries in digital images through dichromatic plane histogram discrepancies. In: Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing, p. 62. ACM (2016)
26. Morrison, K.: How many photos are uploaded to snapchat every second? (2015). <http://www.adweek.com/socialtimes/how-many-photos-are-uploaded-to-snapchat-every-second/621488>
27. Pass, G., Zabih, R., Miller, J.: Comparing images using color coherence vectors. In: Proceedings of the Fourth ACM International Conference on Multimedia, pp. 65–73. ACM (1997)
28. Qureshi, M.A., Deriche, M.: A bibliography of pixel-based blind image forgery detection techniques. *Signal Process. Image Commun.* **39**, 46–74 (2015)
29. Redi, J.A., Taktak, W., Dugelay, J.L.: Digital image forensics: a booklet for beginners. *Multimedia Tools Appl.* **51**(1), 133–162 (2011)
30. Riess, C.: Physics-based and Statistical Features for Image Forensics. Ph.D. thesis, University of Erlangen-Nuremberg (2012)
31. Riess, C., Angelopoulou, E.: Scene illumination as an indicator of image manipulation. *Inf. Hiding* **6387**, 66–80 (2010)
32. Riess, C., Unberth, M., Naderi, F., Pfaller, S., Stamminger, M., Angelopoulou, E.: Handling multiple materials for exposure of digital forgeries using 2-D lighting environments. *Multimedia Tools Appl.*, 1–18 (2016)
33. Rocha, A., Scheirer, W., Boulton, T., Goldenstein, S.: Vision of the unseen: current trends and challenges in digital image and video forensics. *ACM Comput. Surv. (CSUR)* **43**(4), 26 (2011)

34. Shafer, S.A.: Using color to separate reflection components. *Color Res. Appl.* **10**(4), 210–18 (1985)
35. Stehling, R.O., Nascimento, M.A., Falcão, A.X.: A compact and efficient image retrieval approach based on border/interior pixel classification. In: *Proceedings of the eleventh international conference on Information and knowledge management*, pp. 102–09. ACM (2002)
36. Swain, M.J., Ballard, D.H.: Color indexing. *Int. J. Comput. Vis.* **7**(1), 11–32 (1991)
37. Taimori, A., Razzazi, F., Behrad, A., Ahmadi, A., Babaie-Zadeh, M.: A novel forensic image analysis tool for discovering double jpeg compression clues. *Multimedia Tools Appl.*, 1–35 (2016)
38. Tan, R.T., Nishino, K., Ikeuchi, K.: Color constancy through inverse-intensity chromaticity space. *J. Opt. Soci. Am. A* **21**(3), 321–34 (2004)
39. Tralic, D., Zupancic, I., Grgic, S., Grgic, M.: Comofod; new database for copy-move forgery detection. In: *ELMAR, 2013 55th International Symposium*, pp. 49–54, September 2013
40. Van De Weijer, J., Gevers, T., Gijzenij, A.: Edge-based color constancy. *IEEE Trans. Image Process.* **16**(9), 2207–14 (2007)
41. Vidyadharan, D.S., Thampi, S.M.: Detecting spliced face in a group photo using PCA. In: *2015 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, pp. 175–180, November 2015
42. Vidyadharan, D., Thampi, S.: Brightness distribution based image tampering detection. In: *2015 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*, pp. 1–5, February 2015
43. Wu, X., Fang, Z.: Image splicing detection using illuminant color inconsistency. In: *2011 Third International Conference on Multimedia Information Networking and Security (MINES)*, pp. 600–603. IEEE (2011)

A Fast, Block Based, Copy-Move Forgery Detection Approach Using Image Gradient and Modified K-Means

V. Hajhashemi¹ and A. Alavi Gharahbagh²(✉)

¹ Software Department, Bam Pardazesh Tehran Co., Tehran, Iran
Hajhashemi.vahid@yahoo.com

² Department of Electrical and Computer Engineering, Islamic Azad University, Shahrood Branch, Shahrood, Iran
Dramalavi_gharah@yahoo.com

Abstract. In recent years, due to the fast development of digital images, a rapid growth of research interest in the forgery detection in digital images has been happened. One of the most common techniques in creating forged images is copy-move (region duplication) technique. In this paper, a new method for copy-move forgery detection in digital images is proposed. In this paper a region duplication detection technique which utilizes the image gradient is proposed. In the proposed approach, first the gradient of image is divided into overlapped blocks. Using gradient versus other techniques, decreases processing time in feature extraction step.

A fast pre clustering algorithm is another added step to speedup method by dividing search area into some subset. The unknown parameters of proposed method are determined by implementing different conditions on two standard databases. Finally, the performance of the proposed method is compared with some state of art methods and the acceptable accuracy and lower run time of it, is verified.

Keywords: Image forgery · Copy-move · Image gradient · Fast k means · Forgery detection

1 Introduction

In recent years, due to the fast development of image processing techniques and graphics editors, digital images are easily and masterly forged. Image forgery detection have many applications such as forensic and criminal investigation, insurance industry, security systems, social networks, internet, medical imaging etc. One of the most common techniques in creating forged images is copy-move (region duplication) technique. In this method, one or some parts of an image is selected, copied and pasted onto other regions of the same image as shown in Fig. 1.

It is very important in internet or social networks to verify that the images are genuine, so many researches have been done for copy-move forgery detection (CMFD). Existing CMFD techniques can be classified into two major categories: block-based methods and keypoint based methods.

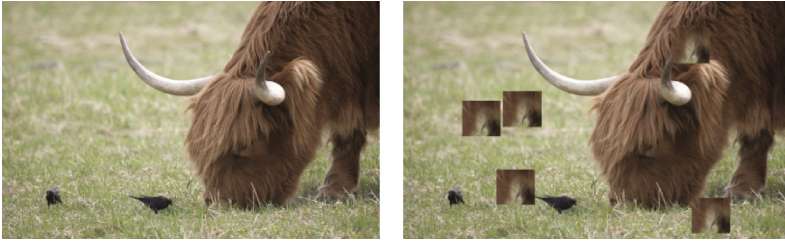


Fig. 1. Original image (right) and forged Image (left)

In block-based CMFD method [1, 2], image is divided into fixed-size overlapping blocks and then similar blocks searched and found. The main differences between these methods are the feature extraction strategy and search area reduction techniques [3]. Because of the large number of blocks, in general, block-based methods are computationally so expensive and the computational burden of feature extraction algorithm and the preprocessing technique for eliminating some blocks from comparison loop is so important [4].

Keypoints-based CMFD methods is approximately different. In these methods at first step, keypoints are detected in image [5]. After this step, similar regions only based on keypoints, searched and found. Obviously keypoints are much less than total overlapped blocks of an image, so keypoints based method are faster than block based methods. The main challenge in these methods are keypoints detection algorithm.

2 Related Work

The main idea in CMFD algorithms is to search and detect similar blocks or regions in forged images. The speed, robustness and accuracy of this detection are some measures used for evaluating effectiveness of different algorithms. considering the scope of the proposed method, in this section only focused on the work that improve copy move forgery detection techniques. [6] suggested a robust approach which use a combination of human vision information and moment invariants features to represent images. In [6] feature vector is a set of color perception and object representation features. [7] used Speed-Up Robust Features (SURF), Histogram Oriented Gradient (HOG) and Scale Invariant Features Transform(SIFT) for CMFD. [8] using SURF similar to [7], but try to improve keypoints detection by an extra step before extracting SURF features. This step uses a single image super resolution (SISR) algorithm. The [8] method showed accurate forgery detection even when the forgery region size is small. [9] suggested mirror-reflection invariant feature transform (MIFT) as an alternative for SIFT. Flipped images are moving images that is generated by a mirror-reversal of original images. [10] is similar to [7, 8] and used SURF feature. For solving the problem of number of keypoints, [10] improve keypoints detection step by particle swarm optimization (PSO). [11, 12] used Discrete Cosine Transform (DCT) for forgery detection. [11] suggested a dynamic threshold to discard large flat regions of the image. Using this threshold, the number of candidate blocks were decreased and the speed of method was

increased. [12] decreased feature vector size in DCT by applying a Zigzag scanning to DCT matrix. In the meantime, in [12] a fast k-means clustering method was utilized to divide blocks to some subsets and reduce the computational burden. In [13, 14] discrete wavelet transform (DWT) was selected for extracting features from image blocks. [13] used stationary wavelet transform and [14] used undecimated dyadic Wavelet Transform. The main difference between [13] and [14] methods are its distance measure that are Euclidean distance in [13] and Canberra distance in [14]. In [15] method a PSO optimization is added to CMFD process to estimate similar regions that are not forged. [16] proposed a key point CMFD method that used Harris corner detection algorithm to find region of interest area (ROI) and eliminate unnoticeable blocks. In [17] a deep learning method is utilized for CMFD. This method did not use any conventional feature extraction method. In deep learning of [17], after patch sampling, by combination of a preprocess including high pass filtering and spatial rich model try to capture artifacts in image. Many other researches has been done in this topic specially in key point based CMFD methods. As a good survey, [18] review some of these works.

All of the aforementioned methods have some limitations. The main challenge in CMFD is computational burden. Specially in high quality images this burden is not acceptable. Key point based methods, try to reduce run time but the accuracy of them is usually lower than block based methods.

In this paper, we propose a system including fast feature extraction process and an extra preprocess step for eliminate some unimportant areas. The other advantage of this step is, its feature participation in final feature vector. Based on this modification the computational cost of total CMFD process is obviously decreased without any accuracy lost. The rest of paper is organized as follows: Sect. 2 explains the related work. Then, the proposed method is clarified in Sect. 3. Comparison with other CMFD algorithms by using a standard dataset is presented in Sect. 4. Finally, conclusion is drawn in Sect. 5.

3 Proposed Method

The details of the proposed algorithm steps are explained in following subsections.

3.1 Image Resizing and Converting to Gray

The proposed algorithm considers the grayscale images. At the first step, if the input image is a color image, it should be transformed to grayscale. In the meantime, the size of the input image directly influences the computation time of CMFD method.

Based on analyzing block size and other conditions, the images larger than 512×512 can be resized to a lower resolution without any decrease in accuracy. For comparing to other methods in similar conditions the resizing step do not apply to images but as an extra step the results of image resizing are given in result section.

3.2 Image Gradient

There are number of ways to extract the unnoticeable pixels from an image. In our CMFD method, an energy value is assigned to each pixel by using gradient energy function that is defined in Eq. (1).

$$e(x, y) = \left| \frac{\partial I}{\partial x} \right| + \left| \frac{\partial I}{\partial y} \right| \quad (1)$$

This value can be easily computed by Sobel masks in both horizontal and vertical directions [19]. The gradient operator is independent of image blocking so the gradient operator can be applied to whole image before blocking. Extracting features from whole image without blocking is so faster than extracting feature vector from overlapping blocks separately. Another important usage of gradient image is discarding unnoticeable blocks in preprocess step. Figure 2 shows original forged image and its gradient.



Fig. 2. CMF image and its gradient

3.3 Image Blocking

In this step, the gradient of input image is divided into overlapping square blocks ($b \times b$ pixels). If the image size is $M \times N$ with overlapping, we have total $(M-b+1) \times (N-b+1)$ blocks in the whole image. For a 128×128 image and block size 8×8 , the total number of blocks is 14641 that with increasing image size this value increase so fast.

3.4 Gradient Threshold

One of the main goals of all CMFD methods is speed of execution. In our method for earning speed, we try to decrease number of blocks before feature extraction step. For eliminating some blocks, the gradient energy used as a fast descriptor. All forged areas based on its nature have discontinuous borders. If total energy of a block be smaller than a specified threshold, block does not include any border and can be eliminated

from next steps and vice versa. All smooth surfaces of image are discarded using this step. For a sample image about 25% of blocks is eliminated in this step that means the total process of our method is about 25% faster than conventional CMFD block-based method. In Fig. 2 the forged region borders are obviously highlight in the gradient image and the low energy of pixels in the smooth surfaces such as meadow is shown by dark areas.

3.5 Feature Extraction

In this stage, singular value decomposition (SVD) is utilized for feature extraction. In fact, the SVD summarizes some properties of a matrix. In the proposed method SVD is applied to each remained block. If the size of the block is $b \times b$, there are at most b singular value for the block, so the proposed method does not need feature reduction step similar to DCT [12] or SIFT features. Some of the singular values is so lower than others and in many matrixes the number of singular values is lower than b (the size of matrix). According to the above reasons, in the proposed method only the 5 first values of SVD are selected as feature vector. If the number of SVD values in a block less than 5, for matching all blocks feature vectors, the remained singular values is supposed 0.

The output of SVD method is sorted descending so the proposed method does not need extra step for sorting each feature vector. In the meantime, the first SVD value computed in this step that is the largest, is used in the next step for clustering blocks into similar groups.

3.6 Modified K-Means and Initial Clustering

The most common method for clustering is k-means. The conventional k-means algorithm is slow for large datasets so in the proposed method a modified accelerated k-means [20] is used for clustering. In conventional k-means most distance calculations are redundant and can be eliminated. In addition, if a point is far away from center, it is not necessary to calculate the exact distance between the point and center, if a point is much closer to one center than other centers, calculating distance is not necessary and this point should be assigned to this center. The accelerated K-Means algorithm [20] avoids unnecessary these unnecessary calculations by applying the triangle inequality and keeping track of lower and upper bounds for distances between points and centers. The upper and lower bounds are usually tight for most points and centers so the updated bounds tend to be tight at the start of next iteration. In fact, the main effectiveness of this method is, reducing the number of distance calculations at the start of each iteration. [20] results showed an about 7% to 300% speedup (based on the number of clusters) in this method in comparison to standard k-means.

In our work this k-means is about 10% faster than standard k-means. It should be noted that k-means is not the main time consuming step of the proposed CMFD process, but surely modification of this process can be accelerated total process. Before finding matched blocks in the final stage, a fast high-performance parallel radix sort

routine for many core GPUs [21] is used for sorting feature vectors in each class. In the proposed method, k-means divided feature vectors into some clusters and the sort routine can be applied to each cluster via a parallel form, so the selected algorithm is fully better than standard quick sort method. The number of clustered is specified in experimental results section.

3.7 Find Matched Blocks

After sorting feature vectors in each cluster, each feature vector only is compared to next feature vector. This means the proposed method only need M comparison in each cluster that M is the number of blocks belong to cluster. For comparing feature vectors, the simple absolute percentage error with a threshold is used. Firstly, the ratio between all feature vectors elements with the same element in the next block is computed. The mathematical explanation of the matching process is given in Eq. 2.

$$Measure = 100 * abs\left(1 - \frac{Feature_vector_{i+1}}{Feature_vector_i}\right) \quad (2)$$

If any of feature vector elements, tend to zero, this element will be discarded from Eq. 2 to avoid error. This measure is applied to each element of feature vectors separately and finally results is compared to threshold.

The threshold value is set to 2 based on analyzing differences between feature vectors in similar regions in noisy and ideal conditions. If this measure in all elements are lower than threshold two blocks are marked as matched or forgery region.

4 Experiment Results

The proposed method was implemented in Matlab 2016a on a computer with CPU core i5-4430, 3.0 GHz with memory 8 GB. The images used in the simulation step, were taken from Christlein et al.'s database [22] and Ng et al.'s database [23].

The properties of images in these two datasets are different. There are 48 tampered images with high resolution in the Christlein et al.'s database. The size of images is about 3000×2300 pixels. The images in Ng et al.'s database are 128×128 pixels gray BMP format image. One of the main parameters in the proposed method is detection time. By choosing these two different datasets, we able to compare detection time of the proposed method with other methods accurately. In order to decrease run time, all images that the sum of height and width of image is more than 1024 are resized to half size of original image, for example a 3000×2300 image is resized to 1500×1150 .

The main parameters in the proposed method were set as: Block size = [4...16], Gradient threshold = 4.68, Minimum block distance between detected forged regions = $2 * \text{Block size}$, measure for finding matched blocks = 2. The block size is

changed between 4 to 16 to test this effect on accuracy and run time of the proposed method. Gradient threshold, Minimum block distance and measure for finding matched blocks were chosen based on the best results (maximize accuracy).

Many metrics have been used in researches for quantifying performance of CMFD methods. The accuracy and robustness [13], Precision and recall [12], True Positive and false positive Rate [9] and finally run time are some of them. In this research we used Precision and recall as a two standard metrics for showing performance of the proposed method and run time for showing processing time. precision (positive predictive value) and recall (sensitivity) are defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4)$$

where True Positive (TP) indicates count of successfully detected forged images. True Negative (TN) is count of normal images successfully labelled by proposed method. False Positive (FP) is the number of normal images detected as forged. False Negative (FN) is number of forged images labelled as normal.

Test (A) Normal CMF Detection (Rectangle region): In this test a normal tampered image without any noise or compression is used for CMFD. The tampered region has fully rectangular form. Figure 3(a, c) and (b, d) show the input forged images and analyzed images respectively. As shown in Fig. 3(b and d) the proposed method has correctly detected copy-move regions in the input tampered images. In the case of Fig. 3(c and d) that the number of forged regions is more than one, only one region did not detect accurately.

Test (B) Normal CMF Detection (irregular region): similar to test A, normal tampered image without any noise or compression is used for CMFD but the tampered area is irregular. Figure 4(a and b) show the input forged image and analyzed image respectively. As shown in Fig. 4 the proposed method has correctly detected copy-move regions in the input tampered images even in irregular form.

Test (C) The unknown parameters of proposed method are number of clusters in initial clustering and the block size. For analyzing these parameters effect on our method, the proposed method applied to 120 images with similar size and the mean of Precision, recall and processing time are computed. Obviously the lower runtime and higher precision and recall value is desirable. The Table 1 shows results. Based on the results of Table 1 the block size 8 and N of clusters 4 are the best values. In the final step the proposed method is compared with some state of art CMFD methods that had similar conditions. For comparing processing time correctly, all methods were implemented on similar hardware. The results are given in Table 2. The precision and recall are given only with one decimal place based on similar works.

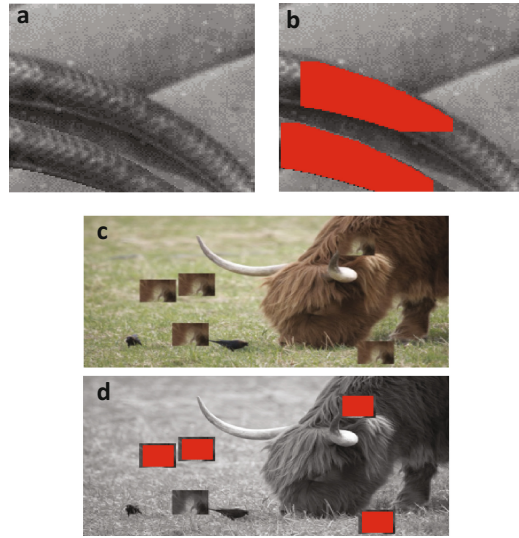


Fig. 3. (a, c) show CMF images and (b, d) the proposed method output

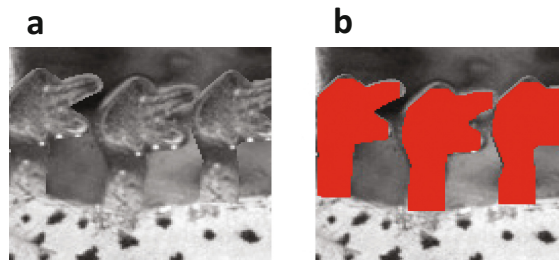


Fig. 4. (a) CMF images and (b) the proposed method output

Table 1. The processing time, precision and recall for different block size and number of clusters.

Block size	N of clusters	Processing time (s)	Precision	Recall
4	4	2.9	87.3	98.1
	8	3.07	88.6	98
	12	3.15	91	98
	16	3.26	94	98
8	4	2.82	99.1	99.4
	8	2.95	98.2	99.2
	12	3.11	97	99.1
	16	3.24	97.3	99
12	4	2.68	96	99.2
	8	2.9	97	99.4
	12	3.22	96	99.1
	16	3.61	94	99

Table 2. The processing time, precision and recall for comparing proposed method to different methods

Ref	Processing time (s)	Precision	Recall
[12]	8.95	99.1	99.1
[14]	10.14	99.1	98.3
[16]	6.64	97.5	99.1
[17]	14.2	95.8	99.1
Proposed	2.82	99.1	98.3

5 Conclusion

In this paper, a new method for CMFD in digital images is proposed. The proposed method decreases processing time in feature extraction step. A fast pre clustering algorithm is another added step to proposed method, to divide search area into some subsets and speedup method. The performance of the proposed method has been analyzed in terms of precision, recall and run time. The unknown parameters of proposed method are determined by implementing different conditions on two standard databases. Finally, the proposed method is compared with some state of art methods and the acceptable accuracy and lower run time of it, is verified. As a future work the proposed method should be modified to identify geometrically or noisy transformed duplicated image regions.

References

1. Bayram, S., Sencar, H.T., Memon, N.: An efficient and robust method for detecting copy-move forgery. In: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, pp. 1053–1056 (2009)
2. Lin, H.J., Wang, C.W., Kao, Y.T.: Fast copy-move forgery detection. *WSEAS Trans. Signal Process.* **5**(5), 188–197 (2009)
3. Zimba, M., Sun, X.: DWT-PCA(EVD) based copy-move image forgery detection. *Int. J. Digit. Content Technol. Appl.* **5**, 19–29 (2011)
4. Hu, J., Zhang, H., Gao, Q., Huang, H.: An improved lexicographical sort algorithm of copy-move forgery detection. In: 2011 Second International Conference on Networking and Distributed Computing, Beijing, pp. 23–27 (2011)
5. Li, L., et al.: Detecting copy-move forgery under affine transforms for image forensics. *Comput. Electr. Eng.* **40**(6), 1951–1962 (2014)
6. Kushol, R., Salekin, M.S., Kabir, M.H., Khan, A.A.: Copy-move forgery detection using color space and moment invariants-based features. In: 2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA), Gold Coast, QLD, pp. 1–6 (2016)
7. Pandey, R.C., Agrawal, R., Singh, S.K., Shukla, K.K.: Passive copy move forgery detection using SURF, HOG and SIFT features. In: Satapathy, S., Biswal, B., Udgata, S., Mandal, J. (eds.) *Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014. Advances in Intelligent Systems and Computing*, vol 327. Springer, Cham (2014)

8. Al-Hammadi, M.M., Emmanuel, S.: Improving SURF based copy-move forgery detection using super resolution. In: 2016 IEEE International Symposium on Multimedia (ISM), San Jose, CA, pp. 341–344 (2016)
9. Agarwal, V., Mane, V.: Reflective SIFT for improving the detection of copy-move image forgery. In: 2016 Second International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), Kolkata, 2016, pp. 84–88
10. Wenchang, S., Fei, Z., Bo, Q., Bin, L.: Improving image copy-move forgery detection with particle swarm optimization techniques. *China Commun.* **13**(1), 139–149 (2016)
11. Moradi-Gharghani, H., Nasri, M.: A new block-based copy-move forgery detection method in digital images. In: 2016 International Conference on Communication and Signal Processing (ICCSF), Melmaruvathur, pp. 1208–1212 (2016)
12. Fadl, S.M., Semary, N.A.: A proposed accelerated image copy-move forgery detection. In: 2014 IEEE Visual Communications and Image Processing Conference, Valletta, pp. 253–257 (2014)
13. Mahmood, T., Nawaz, T., Mehmood, Z., Khan, Z., Shah, M., Ashraf, R.: Forensic analysis of copy-move forgery in digital images using the stationary wavelets. In: 2016 Sixth International Conference on Innovative Computing Technology (INTECH), Dublin, 2016
14. Dixit, R., Naskar, R.: DyWT based copy-move forgery detection with improved detection accuracy. In: 2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN), Noida, pp. 133–138 (2016)
15. Zhao, F., Shi, W., Qin, B., Liang, B.: A copy-move forgery detection scheme with improved Clone Region Estimation. In: 2016 Third International Conference on Trustworthy Systems and their Applications (TSA), Wuhan, pp. 8–16 (2016)
16. Isaac, M.M., Wilsy, M.: A key point based copy-move forgery detection using HOG features. In: 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT), Nagercoil, pp. 1–6 (2016)
17. Rao, Y., Ni, J.: A deep learning approach to detection of splicing and copy-move forgeries in images. In: 2016 IEEE International Workshop on Information Forensics and Security (WIFS), Abu Dhabi, pp. 1–6 (2016)
18. Warbhe, A.D., Dharaskar, R.V., Thakare, V.M.: A survey on keypoint based copy-paste forgery detection techniques. *Proc. Comput. Sci.* **78**, 61–67. ISSN 1877-0509 (2016)
19. Scharf, H.: Optimal second order derivative filter families for transparent motion estimation. In: 2007 15th European Signal Processing Conference, Poznan, pp. 302–306 (2007)
20. Elkan, C.: Using the triangle inequality to accelerate k-means. In: Proceedings of the Twentieth International Conference on Machine Learning (ICML), Washington DC, USA, vol. 3 (2003)
21. Satish, N., Harris, M., Garland, M.: Designing efficient sorting algorithms for manycore GPUs. In: 2009 IEEE International Symposium on Parallel & Distributed Processing, Rome, pp. 1–10 (2009)
22. Christlein, V., Riess, C., Jordan, J., Riess, C., Angelopoulou, E.: An evaluation of popular copy-move forgery detection approaches. *IEEE Trans. Inf. Forensics Secur.* **7**(6), 1841–1854 (2012)
23. Ng, T.T., Hsu, J., Chang, S.F.: Columbia Image Splicing Detection Evaluation Dataset. <http://www.ee.columbia.edu/ln/dvmm/downloads/AuthSplicedDataSet/AuthSplicedDataSet>

OR Operation Based Deterministic Extended Visual Cryptography Using Complementary Cover Images

K. Praveen^(✉), G. Indhu, and M. Sethumadhavan

TIFAC-CORE in Cyber Security, Amrita School of Engineering, Coimbatore,
Amrita Vishwa Vidyapeetham, Amrita University, Coimbatore, India
k_praveen@cb.amrita.edu

Abstract. In visual cryptography, random shares are generated from a binary secret image whereas in extended visual cryptography, meaningful shares are generated using the scheme. There are lots of constructions in visual cryptography and extended visual cryptography for both general and special access structures. Here in this paper, we are proposing an extended visual cryptographic scheme with improved contrast and less pixel expansion for the special access structures (t, k, n) and $(\frac{n}{2} + 1, n)$ where each participant hold single meaningful image as share. The cover images used for creating shares in the proposed schemes are complementary in nature.

Keywords: Secret sharing · Visual cryptography · Extended visual cryptography · Relative contrast · Pixel expansion

1 Introduction

Visual cryptography is an unconditionally secure secret splitting technique used to generate n shares from a binary secret image (SI). During the distribution phase these shares are given to each of the n participants and SI will be visible in the reconstructed image (RI) only during the reconstruction phase when sufficient participants combine their shares. In visual cryptographic scheme (VCS), for reconstruction the Boolean operators used are OR, AND, NOT, XOR instead of complicated computation as in conventional cryptography. The quality of a VCS is quantified using pixel expansion m and contrast $\alpha.m$. A pixel in SI is converted to m sub pixels in all shares. In RI , grey levels of black and white pixel differ by $\alpha.m$. The participants in the qualified (resp. forbidden) set can (resp. cannot) reconstruct SI . VCS are of different types namely $(2, n)$, (k, n) , (t, k, n) , and general access structure. VCS can also be classified into deterministic and probabilistic scheme depending upon the reconstruction of the secret. In deterministic VCS all the black and white areas of SI will be reconstructed in RI , but in probabilistic VCS the correct reconstruction of all pixels in SI are not guaranteed. Deterministic VCS are introduced in papers [1–3]. The perfect black VCS constructions are discussed in paper [4]. Different constructions for deterministic (t, k, n) access structure are given in papers [5–9]. Extended visual cryptographic scheme (EVCS) is another type of VCS where the shares of SI looks like meaningful. In the literature there

are lot of studies on deterministic [10–17] and probabilistic EVCS [18–24]. These deterministic EVCS constructions use perfect black VCS constructions as building blocks.

Here in this paper, we have constructed OR operation based deterministic EVCS for both (t, k, n) and $(\frac{n}{2} + 1, n)$ access structure using complementary cover images with less pixel expansion and more contrast compared to other constructions in the literature. The deterministic constructions for EVCS in the literature uses either distinct [10, 11] or complementary cover images [12] for creating shares. The rest of the sections are organized in the following way. Section 2 shows basic definitions for VCS, EVCS, (t, k, n) and $(\frac{n}{2} + 1, n)$ access structure. Sections 3 and 4 provides the proposed EVCS for (t, k, n) and $(\frac{n}{2} + 1, n)$ access structure. Section 5 addresses some problems of the proposed schemes and its mitigation techniques. Section 6 gives the conclusion.

2 Preliminaries

2.1 VCS [1, 2]

The basis definition of VCS, qualified set, forbidden set, access structure, the relative contrast and security condition for a VCS, the basis matrices used for sharing a secret 1 and 0 pixels of SI are explained in papers [1, 2]. Below shows example basis matrices S^1 (resp. S^0) for sharing a 1 (resp. 0) pixel in SI for a perfect and non perfect black (2,3)-VCS which is having $m = 3$ and $\alpha = \frac{1}{3}$, when the participant set is $P = \{P_1, P_2, P_3\}$, minimal qualified set is $\Gamma_0 = \{\{P_1, P_2\}, \{P_1, P_3\}, \{P_2, P_3\}\}$ and maximal forbidden set is $Z_M = \{\{P_1\}, \{P_2\}, \{P_3\}\}$.

Example 1: Non perfect black construction: $S^0 = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$ and $S^1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.

Perfect black construction: $S^0 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}$ and $S^1 = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}$.

2.2 EVCS for General Access Structures - Ateniese et al. [10]

In the case of EVCS, the share of the participants belonging to set P is meaningful, and is not random-looking nature as in VCS. Let $C_{sc}^{c_1, \dots, c_n}$ where $c_1, c_2, \dots, c_n \in \{0, 1\}$, are the collection of matrices used to encode a c_i pixel in the image COV_i (cover images or meaningful images) where $1 \leq i \leq n$, corresponding to a sc pixel in SI . Hence there will be a collection of 2^n pairs of $(C_0^{c_1, \dots, c_n}, C_1^{c_1, \dots, c_n})$, obtained corresponding to all possible combination of white and black pixels in the n cover images. The condition for contrast of the reconstructed image (α_R) , contrast of the meaningful shares (α_S) and security is given in Definition 3.1 of paper [10]. The construction given in paper [10] uses hyper graph q coloring.

A hyper graph $H = (V, E)$ is a pair where, $E \subseteq 2^V$ members of V are called vertices and members of E are called edges. A q -coloring of H is a function $GC: V \rightarrow \{1, \dots, q\}$, $|\{GC(x) : x \in E\}| \geq 2, |e| \geq 2, e \subseteq E$ for the graph $H = (V, E)$. The chromatic number of H is denoted as $CN(H) \leq |V|$; where $CN(H)$ is the minimum integer q such that a q -coloring of H exists. This construction uses an arbitrary q -coloring of GC on $H = (P, \Gamma_0)$ to encode n cover pixels c_i , to obtain a sc pixel of SI . When OR-ing all rows of the matrix D generated using the EVCS [10] corresponding to the participants in the qualified set $A \in \Gamma_{Qual}$, an all one vector of length q will be generated. The bounds on the value of q are given in Theorem 6.1 of paper [10]. The EVCS for general access structure is constructed using the Boolean matrices of perfect black general access structure construction given in paper [2]. So it is evident from this construction that the collection $C_{sc}^{c_1, \dots, c_n}$ satisfies perfect black property. Let us define em (resp. m) as the pixel expansion of an EVCS (resp. VCS). Then $em = m$ (represents secret information pixels) + q (represents cover image information pixels).

Theorem 1 [10]: For a (k, n) -EVCS, $em = m + \lceil \frac{n}{k-1} \rceil$, where $q = \lceil \frac{n}{k-1} \rceil$.

Example 2: Let $P = \{P_1, P_2, P_3\}$, $\Gamma_0 = \{\{P_1, P_2\}, \{P_1, P_3\}, \{P_2, P_3\}\}$ and $Z_M = \{\{P_1\}, \{P_2\}, \{P_3\}\}$. Let the secret be $SI = [1 \ 0]$. For constructing a $(2, 3)$ -EVCS based on Ateniese *et al.* [10] for the given Γ_0 , we need an arbitrary 3-colouring of GC of $H = (P, \Gamma_0)$. Let $GC(P_1) = 1, GC(P_2) = 2, GC(P_3) = 3$. For sharing a pixel 1 (resp. 0) in SI use the matrix pair S_1^{001} (resp. S_0^{011}) constructed using the three cover images $COV_1 = [0 \ 0], COV_2 = [0 \ 1],$ and $COV_3 = [0 \ 1]$ where

$$S_1^{001} = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \text{ and } S_0^{011} = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

Set of all column permutations of S_1^{001} and S_0^{011} are C_1^{001} and C_0^{011} respectively. The em, m, q, α_{RI} and α_S for this EVCS are 7, 4, 3, $\frac{1}{7}$ and $\frac{1}{7}$ respectively.

2.3 (t, k, n) and $((n/2) + 1, n)$ Access Structure

Let $P = \{P_1, P_2, P_3, \dots, P_t, P_{t+1}, P_{t+2}, \dots, P_n\}$ be the set of participants. A VCS with minimal qualified set $\Gamma_0 = \{A: A \subseteq P, P_1 \in A\}$ and $|A| = k$ is called as (l, k, n) -VCS with P_1 as essential participant. $(1, 3, 4), (1, 4, 5)$ are examples for (l, k, n) -VCS. A VCS with minimal qualified set $\Gamma_0 = \{A: A \subseteq P, \{P_1, P_2, \dots, P_t\} \subseteq A\}$ and $|A| = k$ is called as (t, k, n) -VCS with t -essential participants $P_1, P_2, P_3, \dots, P_t$. $(2, 4, 5)$ access structure is an example for (t, k, n) -VCS for $t = 2$. For $(\frac{n}{2} + 1, n)$ access structure $(2, 2), (3, 4), (4, 6), (5, 8)$ are some examples.

Theorem 2 [5]: The pixel expansion and relative contrast for perfect black (l, k, n) -VCS is $2m$ and $\frac{\alpha}{2}$ respectively, where m and α are pixel expansion and relative contrast for perfect black $(k - 1, n - 1)$ -VCS respectively.

Theorem 3 [6, 7]: The pixel expansion and relative contrast for perfect black (t, k, n) -VCS is $2^t m$ and $\frac{\alpha}{2^t}$ respectively, where m and α are pixel expansion and relative contrast for perfect black $(k - t, n - t)$ -VCS respectively.

3 Proposed (t, k, n) -EVCS

Let COV be the cover image of size $(p \times q)$. Then the complimentary image INV of COV is defined as $INV(g, h) = f(COV)(g, h)$ where $f(x) = \begin{cases} 1 & \text{if } x == 0 \\ 0 & \text{if } x == 1 \end{cases}$, $0 \leq g \leq p - 1$, $0 \leq h \leq q - 1$. So we are using the pair (COV, INV) for generating meaningful shares. When the secret $SI(g, h) == 1$, use the following matrix E^1 for generating shares, $E^1 = \begin{matrix} t & \text{rows} \\ n - t & \text{rows} \end{matrix} \left[\begin{matrix} COV(g, h) \\ INV(g, h) \end{matrix} \middle\| S^1 \right]$. When $SI(g, h) == 0$, use the following matrix E^0 for generating shares, $E^0 = \begin{matrix} t & \text{rows} \\ n - t & \text{rows} \end{matrix} \left[\begin{matrix} COV(g, h) \\ INV(g, h) \end{matrix} \middle\| S^0 \right]$. Also any column permutation of the pair (E^1, E^0) can be used for generating meaningful shares corresponding to each pixel in SI . So E^1, E^0 are matrices of size $n \times (m + 1)$ where m is the pixel expansion of (t, k, n) -VCS [5–7]. In the first column of matrices E^1 and E^0 , first t rows are replaced with the cover pixel $COV(g, h)$ and the last $n - t$ rows are replaced with the inverted cover pixel $INV(g, h)$. Tables 1 and 2 gives the comparison of the proposed scheme with other constructions.

Proof of contrast and security: So in the case of a 0 pixel, when the meaningful shares of essential t participant and any of the $k - t$ participants from the remaining $n - t$ participants stacked (OR operation), the Hamming weight of the resultant vector generated is m , but in the case of a 1 pixel, after stacking the Hamming weight of resultant vector (pixel expansion) obtained is, $em = m + 1$. So the relative contrast for the reconstructed image is calculated as, $\alpha_{RI} = \frac{(m+1)-m}{m+1} = \frac{1}{em}$. The design rationale is, when OR-ing $INV(g, h)$ with $COV(g, h)$ the resultant will be always a bit 1. When looking into the cover pixels $(COV(g, h)$ and $INV(g, h))$ instead of secret information pixels (rows from S^0 and S^1), in the case of E^1 and E^0 matrices all the essential t participants hold $COV(g, h)$ and all the remaining $n-t$ participants hold $INV(g, h)$. Since the qualified set contain k participants (t essential participants + $k - t$ remaining participants), the probability of occurrence of bit 1 during the OR operation of cover pixels will be always 1. Here the relative contrast of the meaningful shares is also maintained as $\alpha_s = \frac{1}{em}$. When the shares of the forbidden set of participants stacked no information can be gained by the participants.

Table 1. Comparison for (1, 3, 4)-EVCS

Scheme	Pixel expansion	α_{RI}
Ateniese et al. [10]	10	0.100
Liu et al. [11]	16	0.062
Wang et al. [13]	16	0.062
Yan et al. [16]	16	0.062
Proposed scheme	7	0.140

Table 2. Comparison for (2, 4, 5)-EVCS

Scheme	Pixel expansion	α_{RI}
Ateniese et al. [10]	18	0.055
Liu et al. [11]	25	0.040
Wang et al. [13]	32	0.031
Yan et al. [16]	32	0.031
Proposed scheme	13	0.076

Example 3: Let the participants are $\{P_1, P_2, P_3, P_4\}$. Let the basis matrices S^0 and S^1 which follows perfect black property obtained for a (1, 3, 4)-VCS [5] are

$$S^0 = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}, S^1 = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 \end{bmatrix}.$$

Following shows generation of meaningful shares based on our (t, k, n) -EVCS. In this example the value of t is 1.

Case 1: When $SI(g, h) == 1$ and $COV(g, h) == 0$, use $E^1 = \left[\begin{array}{c|c} 0 & \\ \hline 1 & \\ 1 & \\ 1 & \end{array} \right] S^1$

Case 2: When $SI(g, h) == 1$ and $COV(g, h) == 1$, use $E^1 = \left[\begin{array}{c|c} 1 & \\ \hline 0 & \\ 0 & \\ 0 & \end{array} \right] S^1$

Case 3: When $SI(g, h) == 0$ and $COV(g, h) == 0$, use $E^0 = \left[\begin{array}{c|c} 0 & \\ \hline 1 & \\ 1 & \\ 1 & \end{array} \right] S^0$

Case 4: When $SI(g, h) == 0$ and $COV(g, h) == 1$, use $E^0 = \left[\begin{array}{c|c} 1 & \\ \hline 0 & \\ 0 & \\ 0 & \end{array} \right] S^0$

For the Γ_{Qual} the contrast obtained is $7 - 6 = 1$, accordingly $\alpha = \frac{1}{7}$ and $em = 7$.

4 Proposed $((n/2) + 1, n)$ -EVCS

Let COV be the cover image of size $(p \times q)$. Then the complimentary image INV of COV is defined as $INV(g, h) = f(COV(g, h))$ where $f(x) = \begin{cases} 1 & \text{if } x == 0 \\ 0 & \text{if } x == 1 \end{cases}, 0 \leq g \leq p - 1, 0 \leq h \leq q - 1$. So we are using the pair (COV, INV) for generating meaningful shares.

When the secret $SI(g, h) = 1$, use the following matrix E^1 for generating shares where $E^1 = \begin{matrix} f & \text{rows} \\ l & \text{rows} \end{matrix} \left[\begin{matrix} COV(g, h) \\ INV(g, h) \end{matrix} \middle| S^1 \right]$. When $SI(g, h) = 0$, use the following matrix E^0 for generating shares, $E^0 = \begin{matrix} f & \text{rows} \\ l & \text{rows} \end{matrix} \left[\begin{matrix} COV(g, h) \\ INV(g, h) \end{matrix} \middle| S^0 \right]$. Any column permutation of the (E^1, E^0) matrices can be used for generating shares corresponding to each pixel in SI . So E^1, E^0 are matrices of size $n \times (m + 1)$ where m is the pixel expansion of perfect black $(\frac{n}{2} + 1, n)$ - VCS [2, 4]. So in the first column of matrices E^1 and E^0 , first $f = \frac{n}{2}$ rows are replaced with the cover pixel $COV(g, h)$ and the last $l = \frac{n}{2}$ rows are replaced with the inverted cover pixel $INV(g, h)$. Table 3 gives the comparison of the proposed scheme with other constructions for (2, 2) access structure. Figure 1 shows the experimental results. Below shows the comparison of the proposed scheme with related works.

- Based on Theorem 1, the pixel expansion of $(\frac{n}{2} + 1, n)$ -EVCS by Ateniese et al. [10] is derived as $m + 2$, where m is the pixel expansion of $(\frac{n}{2} + 1, n)$ -VCS.
- According to EVCS given in papers [13, 14], $em = 2 \times m$.
- EVCS discussed in papers [11, 12], uses a halftone block size of $2 \times 2, 3 \times 3, 4 \times 4$ etc. for encoding m secret information pixels. When the value of m increases the block size also increases. The meaningful shares carried by different participants are distinct in nature.
- For the proposed $(\frac{n}{2} + 1, n)$ -EVCS, $em = m + 1$ and the meaningful shares of participant are complementary in nature.

Proof of contrast and security: So in the case of a 0 secret pixel, when the meaningful shares of any $\frac{n}{2} + 1$ participants stacked (OR operation), the Hamming weight of the resultant vector generated is m , but in the case of a 1 secret pixel, after stacking the Hamming weight of resultant vector is $em = m + 1$. So the relative contrast for the reconstructed image is obtained as, $\alpha_{RI} = \frac{(m+1)-m}{m+1} = \frac{1}{em}$. The design rationale is, when OR-ing $INV(g, h)$ with $COV(g, h)$ the resultant will be always a bit 1. When looking into the cover pixels ($COV(g, h)$ and $INV(g, h)$) instead of secret information pixels (rows from S^0 and S^1), in the case of E^1 and E^0 matrices all the first $f = \frac{n}{2}$ participants hold $COV(g, h)$ and all the remaining $l = \frac{n}{2}$ participants hold $INV(g, h)$. Since the

Table 3. Comparison for (2, 2)-EVCS

Scheme	Pixel expansion	α_{RI}
Ateniese et al. [10]	4	0.25
Liu et al. [11]	3	0.33
Wang et al. [13]	4	0.25
Zhou et al. [12]	4	0.25
Yan et al. [16]	4	0.25
Lu et al. [17]	6	0.16
Proposed scheme	3	0.33

qualified set contains $\frac{n}{2} + 1$ participants, the probability of occurrence of bit 1 during the OR operation of cover pixels will be always 1. Here the relative contrast of the meaningful shares is also maintained as $\alpha_s = \frac{1}{em}$. When the shares of the forbidden set of participants stacked no information can be gained by the participants.

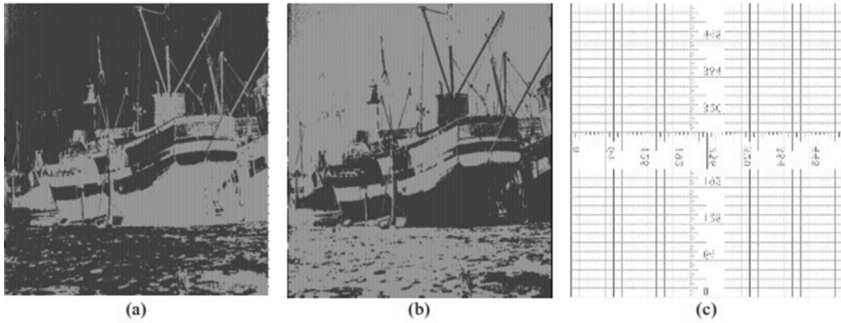


Fig. 1. Experimental results for (2, 2) EVCS (a) share of participant 1 (b) share of participant 2 (c) reconstructed secret

5 Problems with the Proposed Schemes and Its Mitigation

For smaller values of m the proposed schemes outperform other related constructions. But according to the constructions of Wang et al. [13] and Yan et al. [16], the pixel expansion of an EVCS need to be $2m$, where m is the pixel expansion of a VCS. This condition will be helpful to preserve the cover information in the share when the value of m (secret information pixels) increases. So in our scheme, in shares only one bit is assigned for the cover information and the remaining m bits are assigned for secret information pixels. So in the proposed scheme when m increases, the cover pixels will not be highlighted in the shares means the shares look like random. In order to avoid this instead of assigning only one cover pixel, we need to assign r cover pixels to form the pixel expansion as $em = m + r$ instead of $em = m + 1$. The value of r need to be increased gradually till the shares looks meaningful. This is applicable for (t, k, n) and $(\frac{n}{2} + 1, n)$ access structures. So when the value of $r = 2$ the E^1 and E^0 matrices for our (3, 6)-EVCS is shown below for different cases.

Case 1: When $SI(g, h) == 1$ and $COV(g, h) == 0$, use $E^1 = \left[\begin{array}{cc|c} 0 & 0 & \\ 0 & 0 & \\ 0 & 0 & \\ 1 & 1 & \\ 1 & 1 & \\ 1 & 1 & \end{array} \right] S^1$

$$\text{Case 2: When } SI(g, h) == 1 \text{ and } COV(g, h) == 1, \text{ use } E^1 = \left[\begin{array}{cc|c} 1 & 1 & \\ 1 & 1 & \\ 1 & 1 & \\ 0 & 0 & \\ 0 & 0 & \\ 0 & 0 & \end{array} \right] S^1$$

$$\text{Case 3: When } SI(g, h) == 0 \text{ and } COV(g, h) == 0, \text{ use } E^0 = \left[\begin{array}{cc|c} 0 & 0 & \\ 0 & 0 & \\ 0 & 0 & \\ 1 & 1 & \\ 1 & 1 & \\ 1 & 1 & \end{array} \right] S^0$$

$$\text{Case 4: When } SI(g, h) == 0 \text{ and } COV(g, h) == 1, \text{ use } E^0 = \left[\begin{array}{cc|c} 1 & 1 & \\ 1 & 1 & \\ 1 & 1 & \\ 0 & 0 & \\ 0 & 0 & \\ 0 & 0 & \end{array} \right] S^0.$$

6 Conclusion

For the related works given in papers [11, 14–16], all participants use distinct cover images, but in paper [12] all participant use complimentary cover pairs. In this paper construction of EVCS using complimentary cover pairs are proposed for the access structures (t, k, n) and $(\frac{n}{2} + 1, n)$ with a pixel expansion of $em = m + 1$ and relative contrast $\frac{1}{em}$. Compared to related works proposed methods shows better pixel expansion and relative contrast values. In the deterministic OR based EVCS schemes [11, 12, 14–16], all the cover image information pixels q may not be same to the cover pixel c_i in the cover image COV_i corresponding to the participant. But in the proposed scheme, all the cover image information pixels q is same as cover pixel cv (resp. iv) in the cover image COV (resp. inverted cover image INV). So in the proposed scheme when the value, $q = r$ increases the meaningful share look similar to the cover image used. The advantage of the proposed scheme relay in the complimentary pairs of cover images used by the participants.

References

1. Naor, M., Shamir, A.: Visual cryptography. In: EUROCRYPT. Lecture Notes in Computer Science. Springer, Heidelberg, vol. 950, pp. 1–12 (1994)
2. Ateniese, G., Blundo, C., DeSantis, A., Stinson, D.R.: Visual cryptography for general access structures. Inf. Comput. **129**(2), 86–106 (1996)

3. Adhikari, A.: Linear algebraic techniques to construct monochrome visual cryptographic schemes for general access structure and its applications to color images. *Des. Codes Cryptogr.* **73**(3), 865–895 (2013)
4. Blundo, C., Bonis, A.D., Santis, A.D.: Improved schemes for visual cryptography. *Des. Codes Cryptogr.* **24**(3), 255–278 (2001)
5. Arumugam, S., Lakshmanan, R., Nagar, A.K.: On $(k, n)^*$ -visual cryptography scheme. *Des. Codes Cryptogr.* **71**(1), 153–162 (2014)
6. Guo, T., Liu, F., Wu, C.K., et al.: On (k, n) visual cryptography scheme with t essential parties. In: LNCS, vol. 8317, pp. 56–68 (2014)
7. Dutta, S., Rohit, R.S., Adhikari, A.: Constructions and analysis of some efficient $t - (k, n)$ -visual cryptographic schemes using linear algebraic techniques. *Des. Codes Cryptogr.* **80**(1), 165–196 (2016)
8. Praveen, K., Rajeev, K., Sethumadhavan, M.: On the extensions of $(k, n)^*$ -visual cryptographic schemes. In: International Conference on Security in Computer Networks and Distributed Systems, pp. 231–238 (2014)
9. Praveen, K., Sethumadhavan, M., Krishnan, R.: Visual cryptographic schemes using combined Boolean operations. *Journal of Discrete Mathematical Sciences and Cryptography* **20**(2), 413–437 (2017)
10. Ateniese, G., Blundo, C., De Santis, A., Stinson, D.R.: Extended capabilities for visual cryptography. *Theor. Comput. Sci.* **250**(1), 143–161 (2001)
11. Liu, F., Wu, C.: Embedded extended visual cryptography schemes. *IEEE Trans. Inf. Forensics Sec.* **6**(2), 307–322 (2011)
12. Zhou, Z., Arce, G.R., Di Crescenzo, G.: Halftone visual cryptography. *IEEE Trans Image Process.* **15**(8), 2441–2453 (2006)
13. Wang, Z., Arce, G.R., Di Crescenzo, G.: Halftone visual cryptography with error diffusion. *IEEE Trans. Inf. Forensics Sec.* **4**(3), 383–396 (2009)
14. Wang, D.S., Yi, F., Li, X.B.: On general constructions for extended visual cryptographic schemes. *Pattern Recogn.* **42**, 3071–3082 (2009)
15. Yang, C.N., Yang, Y.Y.: New extended visual cryptography schemes with clearer shadow images. *Inf. Sci.* **271**, 246–263 (2014)
16. Yan, X., Wang, S., Niu, X., Yang, C.N.: Halftone visual cryptography with minimum auxiliary black pixels and uniform image quality. *Dig. Signal Process.* **38**, 53–65 (2015)
17. Lu, S., Manchala, D., Ostrovsky, R.: Visual cryptography on graphs. *J. Combin. Optim.* **21**(1), 47–66 (2011)
18. Lee, K.H., Chiu, P.L.: An extended visual cryptography algorithm for general access structures. *IEEE Trans. Inf. Forensics Sec.* **7**(1), 219–229 (2012)
19. Guo, T., Liu, F., Wu, C.K.: k out of k extended visual cryptographic scheme by random grids. *Signal Process.* **94**, 90–101 (2014)
20. Chiu, P.L., Lee, K.H.: User-friendly threshold visual cryptography with complementary cover images. *Signal process.* **108**, 476–488 (2015)
21. Ou, D., Sun, W., Wu, X.: Non-expansible XOR-based visual cryptography scheme with meaningful shares. *Signal Process.* **108**, 604–621 (2015)
22. Yan, X., Wang, S., Niu, X., Yang, C.N.: Generalized random grids-based threshold visual cryptography with meaningful shares. *Signal Process.* **109**, 317–333 (2015)
23. Wang, S., Yan, X., Sang, J., Niu, X.: Meaningful visual secret sharing based on error diffusion and random grids. *Multimed. Tools Appl.* **75**(6), 3353–3373 (2016)
24. Yan, B., Wang, Y.F., Song, L.Y., Yang, H.M.: Size-invariant extended visual cryptography with embedded watermark based on error diffusion. *Multimed. Tools Appl.* **75**(18), 11157–11180 (2016)

Empirical Comparison of Different Key Frame Extraction Approaches with Differential Evolution Based Algorithms

Kevin Thomas Abraham^(✉), Manikandan Ashwin, Darshak Sundar, Tharic Ashoor, and Gurusamy Jeyakumar

Department of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Amrita University, Coimbatore, India

{cb.en.u4cse13129, cb.en.u4cse13109, cb.en.u4cse13114, cb.en.u4cse13169}@cb.students.amrita.edu, g_jeyakumar@cb.amrita.edu

Abstract. Key frame extraction is an integral part of video analytics. The extracted key frames are used for video summarization and information retrieval. There exist many approaches for solving key frame extraction problem in video analytics. The focus of this paper is to extend the strategy of integrating Evolutionary Computing technique with a conventional key frame extraction approach, which is proposed by the authors in their previous work, with two other conventional approaches. The conventional approaches considered in this study are *SSIM* (Structural Similarity Index Method) Method, Entropy Method and Euclidean Distance method. This paper also proposes a new approach for key frame extraction by integrating the Euclidean Distance method with Differential Evolution algorithm. The proposed approach is compared with all the existing approaches by its speed and accuracy. It is found from the comparison that the proposed approach outperforms other approaches. The results and discussion related to this experiment study are presented in this paper.

Keywords: Video analytics · Key frame extraction · Differential evolution · ASSIM values · Entropy difference · Euclidean distance

1 Introduction

In the current information era, information is represented and processed in the forms of multimedia. In general, the multimedia information is suffering with redundancy. Especially, in video processing where numerous frames containing similar information are usually processed. Huge amount of information can be obtained from a video. It can be in terms of shots, scenes and frames. Elimination of redundant information is the issue that is being addressed here. If redundant data is removed, it will help in significantly reducing the amount of information that has to be processed. So, extraction of key frames is the most fundamental step in any video compression or retrieval application. It is important to eliminate the frames with redundant or repetitive information during the extraction process. In the last few years, many key frame extraction

algorithms focusing on live stream videos has been proposed. Key frames represent the prominent features of a video. The extracted key frames must be able to summarize the video without leaving out any key details and it should give the user a complete summary of the video.

The Evolutionary Computing (*EC*) field in computer science, that follows the Darwinian theory of evolution, has many optimization algorithms (popularly known as Evolutionary Algorithms (*EAs*)) in its repository. Finding solutions for a problem using trial and error method is difficult. For such problems *EAs* can be used to search for the solution randomly. The *EAs* have proven their robustness in solving highly complex real world problems. These algorithms have a common structure with the population initialization, mutation, crossover and selection components. These components are used, iteratively, to find a global optimal solution for a given problem starting with a set of randomly generated initial set of solutions. There exist many instances of *EAs* viz Genetic Algorithm, Genetic Programming, Evolutionary Programming, Evolutionary Strategies and Differential Evolution (*DE*). Among these instances the Differential Evolution (*DE*) algorithm is well known in the literature for its simplicity and robustness.

A novel approach to solve the key frame extraction problem by integrating conventional *SSIM* approach with *DE* algorithm is proposed by the authors in [1]. The proposed algorithm was named as *DE_SSIM*. Presently, in this paper, the authors primarily focus on extending the approach presented in [1] by integrating Entropy Difference method and Euclidean Distance method, also, with *DE* algorithm. This integration leads two more algorithms for key frame extraction problem. The algorithms are named as *DE_Entropy* and *DE_Euclidean*, respectively.

The objective of this paper is to strongly validate the novelty that resides in integration of conventional approaches with the Evolutionary Computing algorithms. The performances of these algorithms are compared with their conventional counterpart approaches. The *DE* unified algorithms show high accuracy and robust performance with good customization possibilities. The *DE_Euclidean* method in particular, gives an accurate result as well as a reduced run time when compared with the other two proposed algorithms.

The rest of the paper is organized as follows. In Sect. 2, the works related to our proposed algorithms are presented. In Sect. 3, the structure of *DE* algorithm is discussed. In Sect. 4, the proposed approach is explained with the design and results. Section 5 contains the conclusion remarks followed by the future scope.

2 Related Works

In [2] the author uses entropy value as global and local feature to extract key frames from a video. In this algorithm, they classify the frames in a shot into different bins. Each bin will have a class of frames containing similar objects and background. In this algorithm entropy value is used as a global feature representing the content of the frame. Centre frame from each bin is chosen as one of the key-frame for the shot. Bins with fewer frames i.e., less than twenty frames are neglected to avoid redundant frames. To remove repeated frames, for e.g., a class with a bunch of students writing exam and

in vigilator sitting still results in many redundant key-frames. To eliminate these redundant key frames each key-frame is compared with every other key-frame to find the duplicate or near duplicate frames. Segmented entropy technique is used to remove redundant frames. To measure the dissimilarity between two frames the standard deviation of the difference of segmented entropies of two frames are calculated. If standard deviation is nearing to zero then the two frames are considered as similar and the second frame is eliminated as it is the duplicate frame.

The paper [3] proposes and a new Euclidean distance estimation (*IMED*) for images. This is done by considering the traditional Euclidean distance estimation algorithm and the spatial relationships between pixels. The paper states that by giving a larger coefficient value to nearby pixels of which Euclidean distance is calculated, a robust distance measure is attained. However, *IMED* is applied if three major conditions for metric are satisfied. The metric coefficient for pixels is represented as function of absolute difference between the pixels. The metric coefficient decreases for increase in function dependency. The functional dependency is a universal function and is independent of image or resolution. For any particular pixel given in an image, if the above conditions satisfy the pixel is said to undergo *IMED*.

Along with the existing, commonly known, approaches for solving key frame extraction problem [4–6], a simple algorithmic framework which combine *DE* algorithm with *SSIM* approach was proposed in [1]. The authors have integrated the *SSIM* approach with *DE* algorithm, named the new algorithm as *DE_SSIM*. The results showed that the *DE_SSIM* significantly outperformed *SSIM*. The threshold for the conventional *SSIM* approach was set based on the inputs. However, the *DE_SSIM* algorithm follows a standard evolutionary approach, which follows generation based calculation consisting of parents and children. The number of parents and children are set based on number of frames on input video. The number of key frames is dependent on the number of children. The average of the fitness of the children forms the threshold from which set of key frames that lie below the average are selected. The proposed *DE_SSIM* method was found to result in quicker and more accurate solution when compared to the conventional *SSIM* approach.

On understanding the superiority of the integration proposed in [1], this paper is aiming at proving the superiority further by extending the integration approach to two more conventional key frame extraction approaches. The two other conventional approaches considered are Entropy Difference Method and Euclidean Distance Method.

3 Classical *DE* Algorithm

As it is used in [1], this research work also uses the *DE* for formulating new key frame extraction algorithms. *DE* was popularly used by the research community for solving different optimization problem. As the key frame extraction problem is an optimization problem, *DE* is well suited for solving it. The algorithmic structure of *DE* is shown in Fig. 1 (as it is taken from [7]). As seen from Fig. 1, *DE* performs an evolutionary process with mutation, crossover and selection. This process is done iteratively over a set of candidates present in the population of initial solutions. Each iteration of this

process is meant as a generation. In each iteration, a new child is produced for each of the candidate in the population. This involved performing differential mutation followed by the crossover operation. Then, the fittest candidate among the parent and the child is selected as survivor for the next generation. Thus, at the end of each generation a new population with the survivors is generated. This process will be repeated until meeting a user defined stopping criteria.

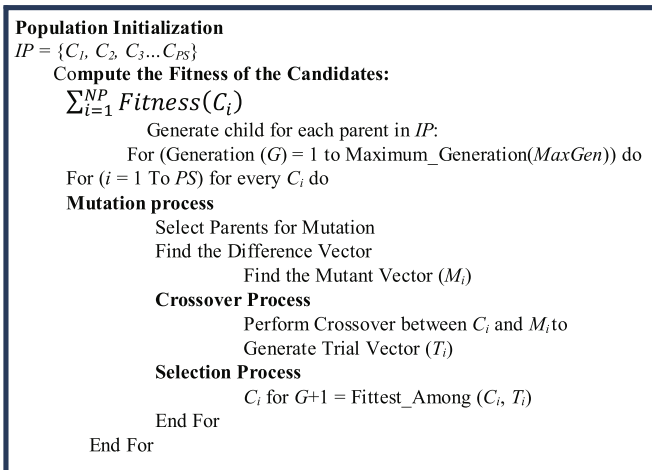


Fig. 1. Algorithmic structure of classical *DE*.

The *DE* algorithm is widely applied for variety of complex optimization problems. The performance *DE* algorithm is still being improved by the researchers [8]. The theoretical investigation of *DE* algorithm is still an open challenge problem [7, 9, 10]. *DE* has only three parameters in its algorithm – Scale Factor (F), Crossover Rate (C_r) and Population Size (NP). Many survey papers are reported in the literature for selecting suitable values for these control parameters [11].

4 Proposed Approach

This paper proposes two algorithms (by exactly following the approach presented in [1]) which incorporates the *DE* algorithm with conventional key frame extraction approaches (a) Entropy Difference method and (b) Euclidean Distance method to solve key frame extraction problem in video analytics effectively. This section discusses the design of the experiment and the results obtained on conducting the experiment.

4.1 Design of Experiment

Classical *DE* algorithm requires certain parameters to be set, these parameters are specific to the problem at hand. Since, the aim of this paper is to further validate the claims made in [1], the control parameters assumed in [1] are used in this experiment as well. However, the fitness function will be different for the two algorithms proposed in this paper. Also, three video samples used in [1] are used again to verify the accuracy of *DE_Entropy* and *DE_Euclidean* algorithms. Frame numbers are input to the algorithm, which in turn will perform *DE* operations and output 10 key frames for each video in this experiment. The fitness Functions used are described below;

- (a) Average Entropy Difference is used as the fitness function in *DE_Entropy* algorithm. Since, it is observed that frames with a higher Entropy Difference result as key frames, the “Selection” operator of the *DE* algorithm will select the vector with a higher Average Entropy Difference value. The mathematical function used to evaluate the fitness of the candidate solutions, for *DE_Entropy* algorithm, is given in Eq. (1).

$$X[MaxF + 1] = \frac{\left(\sum_{i=1}^{MaxF} (Ent(Frame_i) - Ent(Frame_{i+1}))\right)}{(MaxF - 1)} \tag{1}$$

where,

X - a candidate solution (a set of key frame numbers).

MaxF - Maximum number of frames in the key frame sequence, in our experiment it is set as 10.

Frame_i - *ith* frame in a key frame sequence.

Ent(Frame_i) - Entropy value of the *Frame_i*.

- (b) Average Euclidean Distance is used as the fitness function in *DE_Euclidean* algorithm. Since, it is observed that frames with higher Euclidean Distance values result as key frames, the “Selection” operator of the *DE* algorithm will select the vector with a higher Average Euclidean Distance value. The mathematical function used to evaluate the fitness of the candidate solutions, for *DE_Euclidean* algorithm, is given in Eq. (2).

$$X[MaxF + 1] = \frac{\left(\sum_{i=1}^{MaxF} Euc(Frame_i, Frame_{i+1})\right)}{(MaxF - 1)} \tag{2}$$

where,

Euc(Frame_i, Frame_{i+1}) - Euclidean distance between the frames *Frame_i* and *Frame_{i+1}*.

4.2 Results and Discussion

This section discusses the results obtained on executing *DE_Entropy* and *DE_Euclidean*. The results obtained are compared with those of their conventional counterpart approaches (conventional Entropy method and Euclidean Distance method). Tables 1 and 2 represent the results recorded for three different video scenarios, respectively for

DE_Entropy and *DE_Euclidean* algorithm. The results comparing *SSIM* approach with *DE_SSIM* also shown, in Table 3, for reader reference.

Table 1. Comparing *DE_Entropy* and *Entropy* approaches.

Video #	No of frames	<i>Entropy</i> approach		
		Threshold	Key frames	<i>Average Entropy</i> value
Video 1	100	0.018	[27, 34, 36, 47, 49, 92, 94, 95, 96, 97]	0.06
Video 2	409	0.02	[167, 199, 200, 219, 226, 228, 345, 351, 352, 373, 380, 408]	0.05
Video 3	497	0.02	[128, 181, 208, 234, 260, 286, 313, 339, 365, 417, 444, 496]	0.02
Video #	No of frames	Proposed algorithm - <i>DE_Entropy</i>		
		F & C_r	Key frames	<i>Average Entropy</i> value
Video 1	100	0.9 & 0.6	[1, 4, 32, 41, 47, 72, 92, 94, 96, 100]	0.08
Video 2	409	0.9 & 0.6	[137, 146, 176, 200, 211, 224, 249, 274, 317, 354]	0.11
Video 3	497	0.9 & 0.6	[1, 6, 9, 21, 54, 97, 286, 404, 408, 496]	0.05

4.2.1 Entropy Difference Method and *DE_Entropy*

We observe from Table 1 that the Average Entropy value obtained in *DE_Entropy* method is greater than that obtained in the conventional Entropy Difference method. Also, on manual verification of the key frames output by the two approaches, it is observed that the *DE_Entropy* method shows better accuracy. It is worth noting that that the conventional Entropy Difference method required the setting of a threshold value, this value is video specific and varies for the three videos used in this experiment. The threshold values are used as a selection criterion to determine if a particular frame is a key frame or not in the case of the conventional approach. The setting of this parameter is done by trial and error, since it is dependent on the content of the video. However, the *DE_Entropy* algorithm does not require the setting of a threshold value and it will work independent of the content of the video, since no video-specific parameter is to be set.

4.2.2 Euclidean Distance Method and *DE_Euclidean*

We observe from Table 2 that the Average Euclidean Distance value obtained in *DE_Euclidean* is greater than that obtained in the conventional Euclidean Distance method. The frames obtained in both approaches were manually verified and it is observed that the *DE_Euclidean* method shows better accuracy. And in the case of Entropy Difference, a threshold value was to be set for the conventional approach to

Table 2. Comparing *DE_Euclidean* and *Euclidean* approaches

Video #	No of frames	<i>Euclidean Distance</i> approach		
		Threshold	Key frames	<i>Average Euclidean Distance</i> value
Video 1	100	15100	[34, 35, 46, 47, 48, 92, 93, 94, 95, 96]	20490.32
Video 2	409	114000	[208, 209, 210, 362, 363, 364, 365, 366, 367, 368]	13925.33
Video 3	497	16600	[136, 137, 142, 144, 330, 400, 405, 406, 410, 415]	22989.40
Video #	No of frames	Proposed algorithm - <i>DE_Euclidean</i>		
		F & C_r	Key frames	<i>Average Euclidean Distance</i> value
Video 1	100	0.9 & 0.6	[4, 5, 30, 33, 39, 49, 91, 93, 95, 99]	22390.81
Video 2	409	0.9 & 0.6	[20, 44, 151, 178, 210, 214, 348, 363, 372, 409]	22714.28
Video 3	497	0.9 & 0.6	[1, 24, 131, 140, 250, 372, 394, 404, 414, 490]	24405.80

Table 3. Comparing *DE_SSIM* and *SSIM* approaches.

Video #	No of frames	<i>SSIM</i> approach		
		Threshold	Key frames	<i>ASSIM</i> value
Video 1	100	0.87	[27, 29, 31, 33, 35, 37, 45, 47, 92, 97]	0.82
Video 2	409	0.87	[136, 158, 171, 186, 199, 214, 236, 361, 378, 408]	0.73
Video 3	497	0.87	[1, 67, 70, 112, 138, 220, 317, 370, 425, 497]	0.21
Video #	No of frames	Proposed approach - <i>DE_SSIM</i>		
		F & C_r	Key frames	<i>ASSIM</i> value
Video 1	100	0.9 & 0.6	[17, 22, 28, 33, 40, 46, 50, 79, 93, 100]	0.79
Video 2	409	0.9 & 0.6	[136, 157, 171, 187, 198, 214, 236, 361, 377, 409]	0.45
Video 3	497	0.9 & 0.6	[1, 63, 68, 115, 138, 216, 315, 367, 423, 497]	0.11

This table is directly taken from [1]

determine if any given frame is a key frame. This parameter need not be set for the *DE_Euclidean* algorithm and therefore works independent of the content of the video.

Table 4. Execution time comparison

Video #	<i>DE_SSIM</i>	<i>DE_Entropy</i>	<i>DE_Euclidean</i>
Video 1	6 M 29.12 S	3 M 56.87 S	0 M 19.13 S
Video 2	5 M 0.94 S	3 M 14.25 S	0 M 16.83 S
Video 3	4 M 13.35 S	3 M 55.81 S	0 M 25.89 S

Also, it is observed that the running time of the *DE_Euclidean* method was significantly lower than the other *DE* approaches (*DE_SSIM* and *DE_Entropy*). The results comparing the execution time of the proposed algorithms is shown in Table 4. Among the three algorithms *DE_Euclidean* is found as more faster and suitable for further of our research work. As a reference, the resulting key frames on executing *DE_Euclidean* for three different videos are shown Figs. 2, 3 and 4.



Fig. 2. Key frames from Video #1.

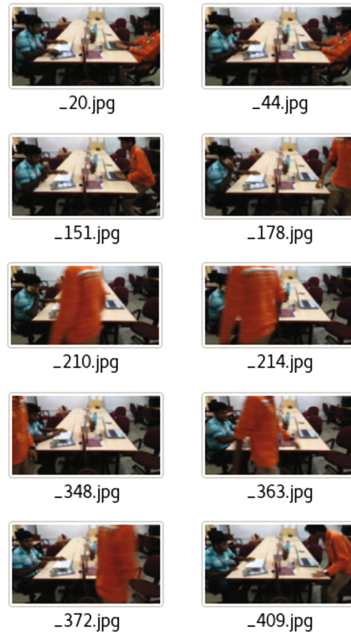


Fig. 3. Key frames from Video #2.



Fig. 4. Key frames from Video #3.

5 Conclusion and Future Work

By directly extending the approach proposed in [1], this paper has proposed two new algorithms for key frame extraction from video. These algorithms combine *DE* with the conventional key frame extraction methods as an attempt to improve the accuracy of key frames extracted from a video. The performance of the proposed algorithms is verified using three experimental videos. Both the algorithms have given more accurate results than their counterpart conventional approaches. The *DE_Euclidean* algorithm was found to have the least execution time amongst the three algorithms (including *DE_SSIM*).

The authors admit here that the observations made are yet to be validated for a real time scenario. This calls the future work that the authors have to carry out. The performance of the proposed algorithm for key frame extraction will be extended for online videos. The future work also includes further reducing the running time by parallelization.

References

1. Thomas, A.K., Ashwin, M., Sundar, D., Ashoor, T., Jeyakumar, G.: An evolutionary computing approach for solving key frame extraction problem in video analytics. In: Proceedings of ICCSP-2017 – International Conference on Communication and Signal Processing (2017)
2. Algur, S.P., Vivek, R.: Video key frame extraction using entropy value as global and local feature (2016). [arXiv:1605.08857](https://arxiv.org/abs/1605.08857) [cs.CV]
3. Wang, L., Zhang, Y., Feng, J.: On the Euclidean distance of images. *IEEE Trans. Pattern. Anal. Mach. Intell.* **27**(8), 1334–1339 (2005)
4. Zheng, R., Yao, C., Jin, H., Zhu, L., Zhang, Q., Deng, W.: Parallel key frame extraction for surveillance video service in a smart city. *PLoS ONE* **10**(8), e0135694 (2015)
5. Sheena, C.V., Narayanan, N.K.: Key frame extraction by analysis of histograms of video frames using statistical videos. *Proc. Comput. Sci.* **70**, 36–40 (2015)
6. Zhang, R., Liu, C.: The key frame extraction algorithm based on the indigenous disturbance variation difference video. *Open Cybern. Syst. J.* **9**, 36–40 (2015)
7. Akhila, M.S., Vidhya, C.R., Jeyakumar, G.: Population diversity measurement methods to analyse the behaviour of differential evolution algorithm. *Int. J. Control Theory Appl.* **8**(5), 1709–1717 (2016)
8. Jeyakumar, G., Velayutham, C.S.: Hybridizing differential evolution variants through heterogeneous mixing in a distributed framework. *Hybrid Soft Comput. Approaches Stud. Comput. Intell.* (Springer) **611**, 107–151 (2015)
9. Raghu, R., Jeyakumar, G.: Mathematical modelling of migration process to measure population diversity of distributed evolutionary algorithms. *Indian J. Sci. Technol.* **9**(31), 1–10 (2016)
10. Raghu, R., Jeyakumar, G.: Empirical analysis on the population diversity of the sub-populations in distributed differential evolution algorithm. *Int. J. Control Theory Appl.* **8**(5), 1809–1816 (2016)
11. Dhanalakshmy, D.M., Pranav, P., Jeyakumar, G.: A survey on adaptation strategies for mutation and crossover rates of differential evolution algorithm. *Int. J. Adv. Sci. Eng. Inform. Technol.* **6**(5), 613–623 (2016)

Breast Cancer Diagnosis and Prognosis Using Machine Learning Techniques

Sunil Suresh Shastri¹(✉), Priyanka C. Nair¹, Deepa Gupta²,
Ravi C. Nayar³, Raghavendra Rao³, and Amritanshu Ram³

¹ Department of Computer Science and Engineering,
Amrita School of Engineering, Bengaluru, Amrita Vishwa Vidyapeetham,
Amrita University, Bengaluru, India

sunilshastri0770@gmail.com, v_priyanka@blr.amrita.edu

² Department of Mathematics, Amrita School of Engineering, Bengaluru,
Amrita Vishwa Vidyapeetham, Amrita University, Bengaluru, India
g_deepa@blr.amrita.edu

³ HealthCare Global Enterprises Ltd (HCG) Hospitals, Bangalore, India
ravi-nayar-ent@hotmail.com, raghav.hcgrf@gmail.com,
dramritanshu.ram@hcgoncology.com

Abstract. Breast cancer is one of the major type of cancer which is the leading cause of death in women. The research work is carried out on the real data of patient records obtained from HealthCare Global Enterprises Ltd (HCG) hospitals. The work analyzes the four major class variables in the dataset, namely death, progression, recurrence and metastasis. The influence of the same 11 predictor variables is explored for each of the class. Various machine algorithms namely Support Vector Machine, Decision Tree, Multi-layer Perceptron and Naive Bayes have been explored for classification of the patient data into various classes. The imbalance in the data is handled using an over sampling technique. The contribution of various attributes in classifying the instances into different classes is also being explored. The model helps in predicting various factors and thus helps in early diagnosis in the breast cancer.

Keywords: Machine learning · Breast cancer · Attribute ranking · Data imbalance

1 Introduction

During the recent years, health industry has been considered as a place where huge amount of data are available including electronic health record (EHR), administrative reports etc. This massive data is not utilized well. Various machine learning algorithms are applied on data to study the hidden patterns in the data and helps in various processes like prediction. ML techniques have been used in health care to predict various diseases. These techniques are also explored to predict the treatment, its outcome, various factors affecting a health condition and many other things interesting to the health care providers [1]. Thus it can help as an aid to the doctors in providing a better treatment for a patient presenting with various diseases. Cancer is a group of diseases which involve

abnormal growth of cells which later may spread to various parts of the body. When the global population is considered, cancer is the second major reason for death. There have been around 8.8 million deaths in the year 2015 due to cancer [2].

There are different types of cancers like Lung cancer, Liver cancer, Stomach cancer, Breast cancer and many more. Breast cancer is one of the most frequently occurring types of cancer, especially in the female population. The cancer needs to be diagnosed at its early stage, before it spreads to the other parts of the body, so that the chances of the patient survival are high [3].

So it is important to have an early diagnosis in case of patient who has a cancer. The prognosis of a cancer patient can be improved if various predictions can be made, to know whether there is chance of recurrence, progression etc. from given set of patient details, with the help of a prediction model using ML techniques.

The various researches works which have been already done in breast cancer focus on predicting one factor (like survival rate/recurrence) at a time. Different work focus on different predictor variables while building the prediction model. In some works, clinical parameters are used whereas in others features extracted from images are used. In this work multiple class variables are predicted using the same features/attributes. The health care data usually has more number of patients who does not have a particular disease condition and a few numbers of patients who have that condition. This is treated to be an imbalance in data if it is 10–20%. In many of the existing research, SEER data set (publicly available well studied data set for breast cancer data) is used, which has millions of records [5]. So many research works using this data set tackles the issue of imbalance by eliminating few records from the majority set. But in the proposed work real data from HCG hospitals, Bangalore has been used. The data is small and hence eliminating records is not a possible solution to avoid imbalance. The proposed work applies an over sampling technique called SMOTE to handle the imbalance in the dataset. SMOTE increases the number of records for the class with less data and keeps the data belonging to the class, which has sufficient data, unchanged. The influences of predictor variables on the different class variables are studied. Ranker algorithm has been applied on the data to know rank the attributes before classification of the data. Simple and well established classification algorithms suitable for medical data are decision tree, Multi-perceptron, Support Vector Machine and Naive Bayes. These algorithms have been applied on the data for prediction. As there are many missing values in the data set, some data cleaning has been performed on the data. The content of the paper is organized into various sections as follows. The related works in this research area is discussed in Sect. 2. Section 3 describes the data set on which the work is experimented. Proposed methodology is discussed in Sect. 4. The various experiments carried out are explained in the Sect. 5. The results and analysis is described the Sects. 6 and 7 discusses the conclusion and future work respectively.

2 Related Works

Over the past decades data in health care is in the form of electronic health record and is growing at a very rapid pace. Data in health care are present in different formats like narrative text data, numerical data, recorded signals or images. Discharge summary

would be in the form of narrative text, radiology uses images, ECG is the form of signals and a large chunk of data is available as numerical data which could be the lab reports, vital signs etc. Works have been done focusing on these multiple forms of health care data. The features are extracted from the image, text or signal data using techniques like image processing, text processing, signal processing and as a next step ML techniques are explored on them to develop different models which can help in decision making. The models study the existing data available for any disease, find and analyze hidden patterns in them. Thus it assists the health care providers in predicting a patient's health condition who is presenting with certain symptoms and clinical conditions. There are several works that apply these kinds of models which aims at predicting different kinds of diseases like cardiovascular diseases, Alzheimer, diabetes, cancer and many other diseases. Study needs to be performed to analyze such huge amount of data.

ML techniques are applied on clinical and claims data to study the correlation between various chronic diseases and the related diagnostic tests conducted [6]. These algorithms can be also applied to study about the effects of a particular disease on some other disease. Effect of diabetes and ischemic heart disease on other diseases (using ICD9 codes) is studied by application of machine learning algorithms [7]. The relationships between different attributes of cardio vascular disease are explored using association mining which is an ML technique [8]. Such various studies are done in health care with the help of different pattern recognition algorithms to improve the diagnosis, treatment etc. in the medical field [1].

In healthcare a major research is on the cancer as it is one of the main reasons of mortality in the current century. Classification algorithms like Artificial Neural Networks (ANNs), Bayesian Networks (BNs), Support Vector Machines (SVMs) and Decision Trees (DTs) have been widely applied in cancer research for prediction of cancer susceptibility, recurrence and survival [9]. Prediction model is developed to predict the survival rate of oral cancer [10]. The work focus on details of 1024 patients with 19 predictor variables which include history of addiction, site, tumor size, staging, surgery, radiotherapy, chemotherapy. Target variable is survival which takes binary value alive or dead. Data is not imbalanced as it has 32% records of patients who have survived and 68% who have died within 5 years of treatment. Importance of predictor variables are also discussed in the work. Three predictive models Single Tree, Decision Tree Forest and Tree Boost are developed and it is observed that Tree Boost model performs well compared to other two models on the oral cancer data. A study is done which aims to use clinical information of nearly 500 lung cancer patients to replace pathology report, so that surgery to collect pathologic information can be avoided [11]. Correlation between clinical information and pathology report is explored using association mining.

Among the various cancers, breast cancer is a very commonly occurring type of cancer [12]. Various researches have been done on breast cancer data using various machine learning algorithms. A prognostic model is developed based on support vector machine (SVM) to predict the recurrence of breast cancer [13]. The model's performance is compared with previously established models. The model helps the doctors and the patients regarding the breast cancer treatment. A work is conducted to analyze the performance of three machine learning algorithms Decision Tree, Support Vector Machine, and Artificial Neural Network on prediction of the breast cancer recurrence [14]. The dataset consist of the 1189 records with 22 predictor variables and 1 outcome

variable. Total number of records reduced to 547 as they have been removed due to missing values. As Age, site of the cancer, tumor size, ER, PR, Her2, Chemotherapy, Radiotherapy etc. were few main independent variables used in the study. There were 117 recurrence cases and 430 non recurrent cases in the dataset and hence no sampling techniques are explored as the data is a balanced dataset. According to the analysis from the work SVM exhibits better performance than DT and ANN in all the parameters of sensitivity, specificity and accuracy. Most of the works use ML techniques for prediction and the performance is evaluated based on accuracy. But stability of the performance of the model when the parameters are varied need to be considered [15]. A stable model will actually help the health care specialists who have little knowledge about the model tuning. The work compares SVM, ANN and semi-supervised learning models on predicting breast cancer survivability and finds semi-supervised learning model to be a stable robust model. The medical data usually has an imbalance, which means the number of instance in various classes may vary largely.

Prediction would be skewed towards the majority class in studies dealing with such imbalanced dataset. Hence research is carried to study various sampling techniques like SMOTE, TOMEK and combined sampling in combination with SVM classification algorithm on few medical datasets available in UCI repository [16]. The work shows that combine sampling outperforms the other two. But when the number of records in the minority class is less than 10%, it will not perform better than the other two algorithms.

Various research works focus on exploring different classification techniques on the data where a single problem is addressed. While a work predicts recurrence of cancer, another work attempts to predict survival rate. Different features are considered for different work. So there has been no model which tries to predict multiple factors considering the same attributes. In the proposed work, same attributes are used to predict various response variables which are death, recurrence of cancer, metastasis and progression of cancer. In various works, using breast cancer data, either balanced data is used or the imbalance is less (minority class is almost 30%). In some cases imbalance is addressed by removing the instances from the majority class, which is called under sampling. This may not be suitable in the proposed work as the numbers of instances are less. Other techniques like combine sampling works well only in the cases where the minority class is more than 10% [16]. In proposed work in some cases the minority class is only around 10% and hence an over sampling technique called SMOTE is applied. The attributes are ranked using Ranker algorithm so that importance of each attribute in classifying the data to the particular class can be understood.

3 Data Source

The data is provided from HCG hospitals by maintaining the anonymity of the patients. The data has records of clinical trials of 1595 breast cancer patients which consist of different medical parameters of which 11 are the predictor variables and 4 are class (response) variables. The 4 class variables are death, metastasis, progression and recurrence. Death is an attribute which is considered to be the class variable, which explains whether the patients have died within 5 years of treatment. Metastasis takes a binary value (0 or 1) based on whether the cancer has spread to the other parts of the

body. Progression means the process of developing gradually towards a more advanced state. Recurrence means the return of the cancer after certain period of time, the value of this attribute explains whether the cancer is likely to occur again. All the class variables take binary value as 0 or 1. The detailed description of the variables and their values for the predictor and response variables are explained in Table 1. The Sect. 4 describes the proposed methodology in detail.

Table 1. Data description of predictor variables and response variables

Attributes	Values	Attributes	Values
Age Category	0 (age below 50) & 1(age above 50)	Progesterone Receptor	1(Progesterone receptor positive) 0(Progesterone receptor negative)
Cancer Type	1(Left), 2(Right) & 3(Bilateral)	(Her2)	1(Her2is present) & 0(Her2 is absent)
Grade	1(well differentiated), 2(moderately differentiated), 3 and 4 (poorly differentiated)	Hormone Receptor Status code	1(ER and PR is positive& Her2 is negative) 2(ER and PR is absent & Her2 is present) 3(ER,PR & Her2 are absent), 4(All three hormone receptors positive)
Stage	1(Tumor size max up till 2cm) 2(Tumor between 2 and 5cm), 3(Tumor size larger than 5cm & have spread to lymph nodes) 4(Tumor have spread to nearby lymph nodes)	Treatment	1(Surgery plus some other treatment is performed), ,2(Chemotherapy, radio therapy and surgery is performed) 3Chemotherapy, radio therapy performed), 4(Only surgery is performed) 5(Only Chemotherapy is performed) ,6(Only radio therapy is performed)
Estrogen-Receptor(ER)	1(Estrogen receptor positive) 0(Estrogen receptor negative)	Surgery	1(Surgery is performed), 0(Surgery is not performed)
Death class	1 (The patient have died within 5 years of the treatment) & 0(The patients have survived for within 5 years of the treatment)	Radio Therapy	1(Radiotherapy is performed), 0(Radiotherapy is not performed)
Metastasis class	1(Cancer has spread from breast to other organs) 0(Cancer has not spread from breast to other organs)	Recurrence class	1(The return of the cancer after certain period of time) 0(No, return of the cancer after certain period of time)
Progression class	1(Cancer progress to advance state in the breast) 0(Cancer does not develop to advance state) 0 (Not developing towards the advance state)		

4 Proposed Method

The proposed methodology is based on five major stages which are data pre-processing stage, sampling stage, classification stage and evaluation stage as shown in Fig. 1. The detailed descriptions of each of these stages are as follows.

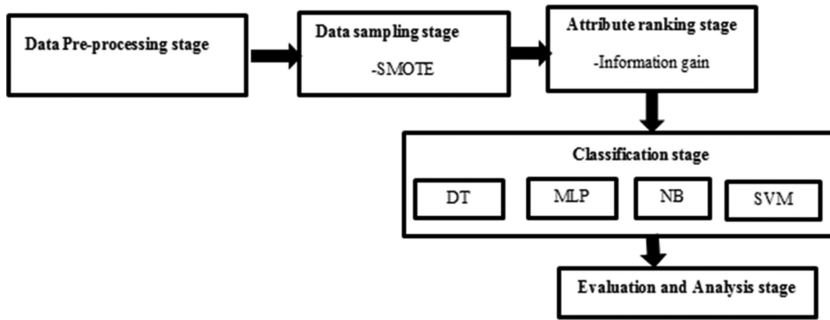


Fig. 1. Schematic diagram of proposed approach

4.1 Data Pre-processing Stage

Any health care data which contains patient records would have values for certain attributes missing. The reason for this would be either few tests may not be performed on the patient so the values for those are not known, or it may be an error in the data entry. Missing values in the data set needs to be removed before application of the proposed model.

4.2 Data Sampling Stage

Usually data in medical field would have imbalance in data. For example number of patients without a disease may be much more than number of patients who have that disease. In the dataset used in this work, all patients are suffering from cancer. Here the number of patients without a condition is large when compared to number of patients with that condition. This kind of imbalance need to handle by sampling technique. Synthetic Minority Over-sampling Technique (SMOTE) is an over sampling method where minority class is increased, by random replicating the data set of the minority class while the majority class is not changed [17]. Each minority class sample is taken and synthetic data are created a long line segments that join the k minority neighbors of the class. Neighbors from k nearest neighbors are chosen depending on the amount of oversampling needed.

4.3 Attribute Ranking Stage

Ranking of the attribute is performed with the amount of the information in a variable, which is information gain, each variable will be given a rank based upon the percentage

of the influence it has on the class variables [18]. Information gain of an attribute is reduction in entropy due to this attribute. Entropy is the level of impurity of a dataset D and is calculated as

$$E(D) = -p \log_2 p - s \log_2 s \quad (1)$$

In the dataset considered in this work, if the recurrence class is considered, p is the ratio of number of recurring cases to total number of records and s is the ratio of number of non-recurring cases to total number of records.

$$Gain(D, V) = E(D) - \sum (D_v/D) * E(D_v) \quad (2)$$

where D_v refer to records having multiple value for an attribute. Here in this work if the attribute Cancer type is considered it takes three values 1, 2 and 3.

4.4 Classification Stage

Simple classifiers with good performance which are Decision Tree (J48), MLP, Naive Bayes and Support Vector Machine (SVM) are applied in the proposed model. These techniques are applied on the dataset to classify data into different classes. The proposed work deals with binary classes, there are different kind of classifiers when applied on the different types on the data will produce similar results. Decision tree (J48) [19] builds the decision tree from the dataset, the roots and various nodes of the tree is decided by information gain, at each node of the tree the algorithm choose the next variable to be the root node on the basic of the information gain, the leaf are the class variable in this algorithm. Multilayer perceptron (MLP) [20] is and a feed forward Artificial Neural Network (ANN) which consists of the multilayer nodes in between where each layer of the nodes are fully connected to the each other MLP maps the set of the input data to the set of the output data which are appropriate. Naive Bayes [21] is the classifier which uses the probability concept which is based on the Bayes theorem, this classifier is highly scalable; by evaluating the closed form expression maximum-likelihood training can be done. Support Vector Machine (SVM) [22] is also a classification algorithm which tries to construct a hyper plane which maximally separates data of two classes. After classification, k-fold cross validation is applied to avoid the problem of over-fitting. The data is partitioned into training set (on which the model is applied) and testing set (on which validation is done) and this is repeated k times in k fold cross validation.

4.5 Model Evaluation Stage

Performance of the model is evaluated using the measure: accuracy, specificity, and sensitivity. Confusion matrix is a table used in to assess the performance of a classification model on the test data set is as shown in the Table 2.

Table 2. Confusion matrix which is used for the model evaluation

Actual	Positive	Negative
Positive	TP	FN
Negative	FP	TN

True Positive (TP) indicates the number of instances the model has predicted presence of a disease (positive) and disease is actually present. False Positive (FP) gives the count of records predicted by the model to have presence of a disease which is actually not present. The number of records predicted by the model as not having the disease but actually but are actually having the disease is called False Negative (FN). The number of instances which are actually not having the disease and the model also predicts as not having the disease is True Negative (TN). ROC [23] (Receiver Operating Characteristic) curve, is a graphical representation which shows the performance of a binary classification model. The curve is constructed by plotting the Sensitivity against (1-Specificity). ROC curve can be analyzed to understand the performance different classifiers for binary classes. Sensitivity, Specificity, Accuracy are calculated as given in the Eqs. (2), (3) and (4) respectively: Attributes have been ranked and analyzed to understand which attributes are contributing towards each of the classes.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (4)$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (5)$$

5 Experimental Setup

The various machine learning algorithms are explored on the dataset and the various experiments are conducted using the publicly available machine learning tool WEKA (Waikato Environment for Knowledge Analysis) version 3.8.0.

Table 3 summarizes the missing value for each attribute in the dataset. This is handled in the pre-processing stage of the model. Instances, where data for at least 5 attributes are missing, have been removed. The rest of the missing values in the various attributes have been filled by replacing with mode of that attribute in that particular class. i.e. When a data is missing in attribute 1 of a record which belongs to class 1, then it has been replaced by the mode of the attribute 1 in class 1. Thus all missing values in the data set have been removed and the data is cleaned. After cleaning the percentage of the missing value is zero.

Table 3. Summary of percentage of missing values in the dataset

Attributes	Missing values in percentage	Attributes	Missing values in percentage
Age category	0	Progesterone receptor	25
Cancer type	0	(Her2)	26
Grade	43	Status code	0
Stage	0	Treatment	28
Estrogen receptor (ER)	25	Surgery	1
Death class	21	Radio therapy	20

The number of instances has reduced after the data cleaning is performed for each class as shown in Table 4. When the number of instances in one class is very large compared to the number of instances in other class, the data set is said to be imbalanced. Classification algorithm will not classify the minority class (the class with less number of instances) appropriately if there is an imbalance in the dataset. This work has made use of an over sampling technique called SMOTE (Synthetic Minority Over-sampling Technique). Numbers of records that belong to each class are reduced after the data cleaning process and is shown in Table 4. After applying the over sampling technique SMOTE on the data for each class, the number of instances in minority class are increased. Various percentage of SMOTE is tried on the data and good performance was shown by the model when 500% SMOTE is applied on the data. The number of instances in majority class would not be changed. The number of instances in both majority and minority class (class 1 in this work) are also shown in the Table 4. The different classifiers are applied with default parameter settings of WEKA as explained in Table 4. The proposed model is compared with base model, in which the classifiers are directly applied on the raw data.

All attributes are ranked using Ranker algorithm. The parameter used is Information Gain. For this research, 10-cross validation technique have been applied, on the dataset to avoid the over fitting problem.

Table 4. Data statistics after preprocessing and sampling.

Class	Death		Metastasis		Recurrence		Progression	
	0	1	0	1	0	1	1	0
Raw data	1412	134	1473	73	1515	31	1374	172
Data after cleaning	1096	77	1138	54	1153	26	1038	135
Data after application of sampling technique	1096	462	1138	24	1153	56	1038	810

Table 5. Parameter settings for the classifiers used in the proposed method.

Classifiers	Basic parameters
Decision tree (J48)	Batch size = 100, BinarySplits = False, Confidence factor = 0.25, NumFolds = 3, Reduce error pruning = False, Seed = 1, Unpruned = False, Use Laplace = False, UseMDLcorrection = True
Multi-layer perceptron (MLP)	Hidden layers = a; learning rate = 0.3; momentum = 0.2; nominal to binary filter = True; normalize attributes = True; normalizeNumericClass = True; reset = True; Seed = 0; trainingtime = 300; validation Threshold = 20
Naïve Bayes	Use Kernel estimator = False; use supervised discretization = False
SVM	Build logistic models = False; c = 1.0; checks turned off = False; epsilon = 1.0E-12; Kernel = Polykernel; num of folds = 1; randomseed = 1; tolerance parameter = 0.001

6 Experimental Results and Analysis

Classifiers are applied on the raw data which is not cleaned and balanced. The performance is shown using ROC curve and a table which depicts the values of accuracy (in percentage), sensitivity and specificity. Various lines in the ROC curve correspond to the results of different classifiers. Multiple classifiers exhibit similar performance which leads to over lapping of curves in the figures. This is compared with the proposed model which has the classifiers applied on cleaned and balanced data. The performance of this model is also shown using ROC curve and table. The comparisons are made and analysis is performed in the following subsections for each of the class. In each figure the multiple curves denotes performance of different classifiers. As the curve moves closer to left-hand border and then the top border of the ROC space, the model is considered to be a model with good performance. When the curve reaches close to the 45° diagonal of ROC spaces, the model is considered to be less accurate (Table 5).

6.1 Experimental Results and Analysis for Death Class

Different classifiers are applied on the base model and the proposed model for the death class and their performance is compared in the ROC curves as shown in Figs. 2 and 3. The Table 6 is the results obtained by applying the various classifiers on base model and the proposed model.

There is a significant improvement when the classification algorithms are employed on the proposed model as compared to the results obtained from the raw data. All the classifiers i.e. DT, MLP, NB, and SVM perform well in terms of accuracy, but MLP classifier which performs best on the proposed model in terms of sensitivity for death class.

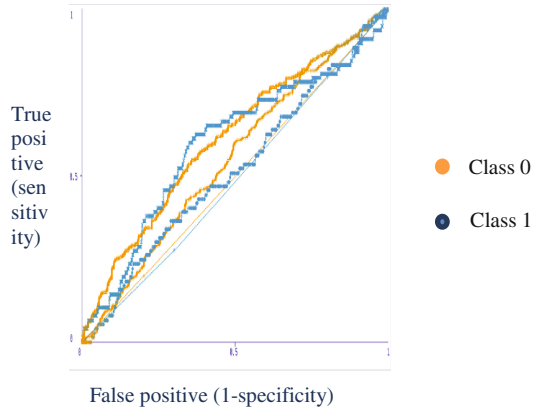


Fig. 2. ROC curve for death class with unprocessed data

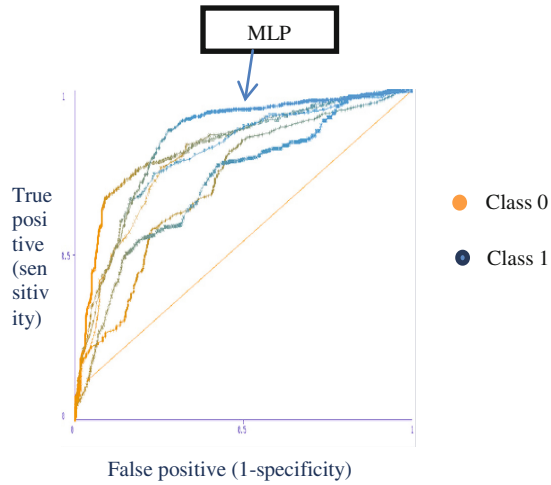


Fig. 3. ROC curve for death class with processed data

Table 6. Class death with unprocessed dataset (base dataset) and proposed dataset

	Base model				Proposed model			
	DT	MLP	NB	SVM	DT	MLP	NB	SVM
Accuracy (%)	92.0	88.0	90.0	93.0	78.0	77.0	74.0	71.0
Specificity	1.0	0.9	0.9	0.9	0.8	0.8	0.8	0.9
Sensitivity	0.0	0.0	0.0	0.0	0.6	0.6	0.5	0.1

6.2 Experimental Results and Analysis for Metastasis Class

The base model is where the classifiers are applied on the raw data of the metastasis class and the performance is compared with the performance of the classification techniques on the proposed model as depicted by the ROC curves in Figs. 4 and 5. In Fig. 4 which represents the base model the curves are near to the diagonal line which presents that the performance is poor. But in the proposed model the curves are moving towards left border and towards top border which indicates a good performance. Table 7 gives a comparison in the values for accuracy, specificity and sensitivity of the base model and proposed model. It is very clearly understood the proposed model gives good performance as compared to the base model. In this class also the classifier which performs well is MLP.

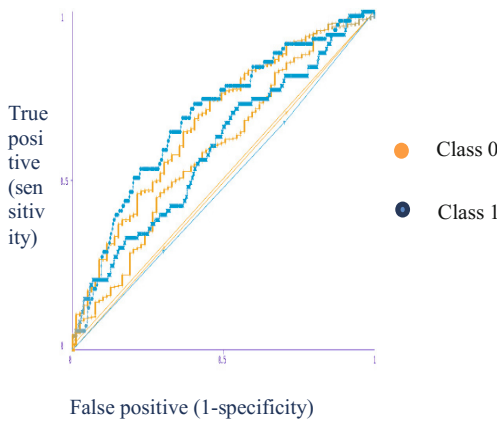


Fig. 4. ROC curve for metastasis class with unprocessed data

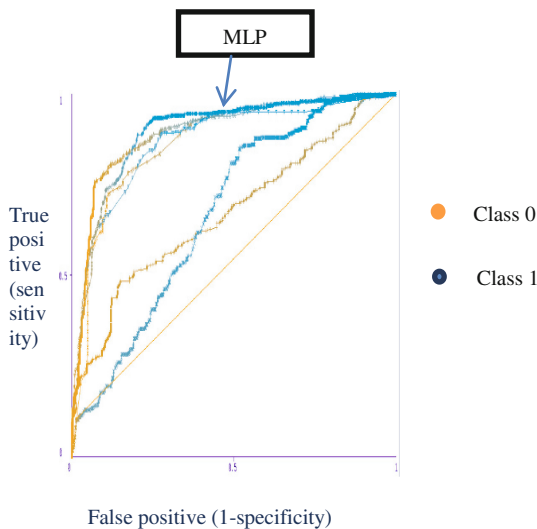


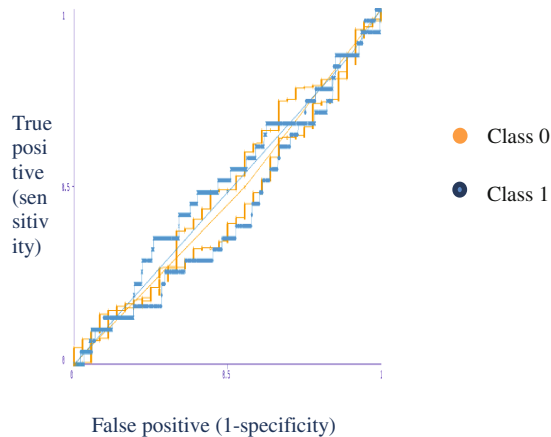
Fig. 5. ROC curve for metastasis class with processed data

Table 7. Class metastasis with unprocessed dataset (base dataset) and proposed dataset

	Base model				Proposed model			
	DT	MLP	NB	SVM	DT	MLP	NB	SVM
Accuracy (%)	95.0	93.0	94.0	95.0	85.6	86.0	78.3	78.8
Specificity	1.0	0.9	0.9	0.9	0.9	0.9	0.9	0.9
Sensitivity	0.0	0.0	0.0	0.0	0.6	0.7	0.1	0.1

6.3 Experimental Results and Analysis for Recurrence Class

Analysis is performed after applying various classification algorithms on the data belonging to recurrence class. Figure 6 shows the results obtained by applying the various classifiers on the raw data. From the figure it is clear that the base model performance is poor as the curves are near to the 45° line. Few of the curves are even below the line. Figure 7 depicts the result obtained by proposed model. Table 8 indicates the results obtained by both base. Model and the proposed model by applying different classification algorithms. Significant improvement can be viewed in ROC curve as well as in the table, after the proposed model has been applied. Though all models are performing well in terms of accuracy, since sensitivity is an important parameter to be considered in health care data, MLP can be treated as the best classification technique for recurrence class.

**Fig. 6.** ROC curve for recurrence class unprocessed data

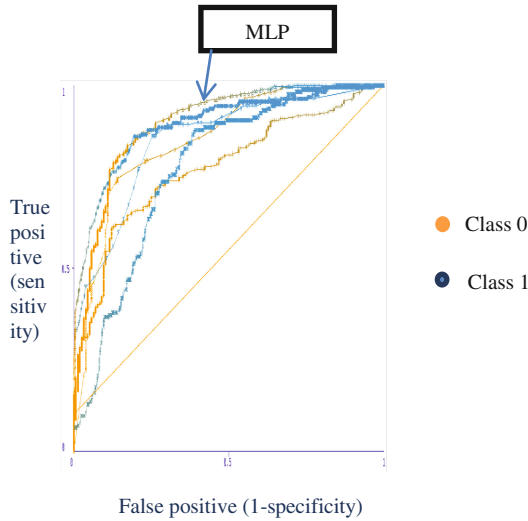


Fig. 7. ROC curve for recurrence class processed data

Table 8. Class recurrence result with unprocessed dataset (base model) and proposed model

	Base model				Proposed model			
	DT	MLP	NB	SVM	DT	MLP	NB	SVM
Accuracy (%)	97.9	97.8	97.8	97.9	91.2	91.2	88.5	88.9
Specificity	1.0	0.9	0.9	1.0	0.9	0.9	0.9	0.9
Sensitivity	0.0	0.0	0.0	0.0	0.3	0.4	0.0	0.1

6.4 Experimental Results and Analysis for Progression Class

Experiment is done on the data which belongs to class progression. Figure 8 show the results obtained by applying the various classifiers algorithms on the raw data and Table 9 shows the result obtained for both the base model and the proposed model by applying the various classification algorithms. The Fig. 9 shows the results of the dataset which processed through the proposed method. Comparing both the figures and table it can be concluded that the proposed method have shown a significant improvements. In Table 9 under the proposed method classifiers decision tree have shown good accuracy but decision tree has performed better with 81.71% of accuracy and sensitivity of 0.84.

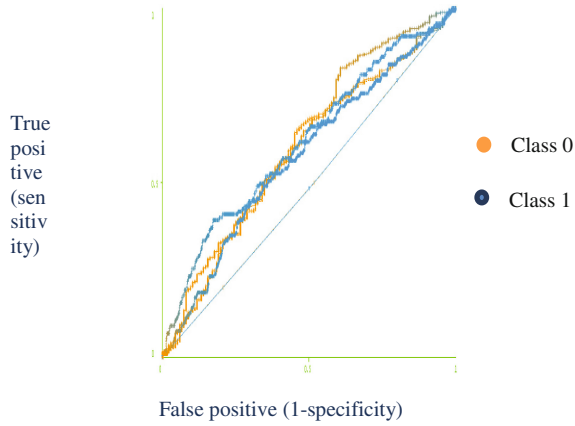


Fig. 8. ROC curve for progression class unprocessed data

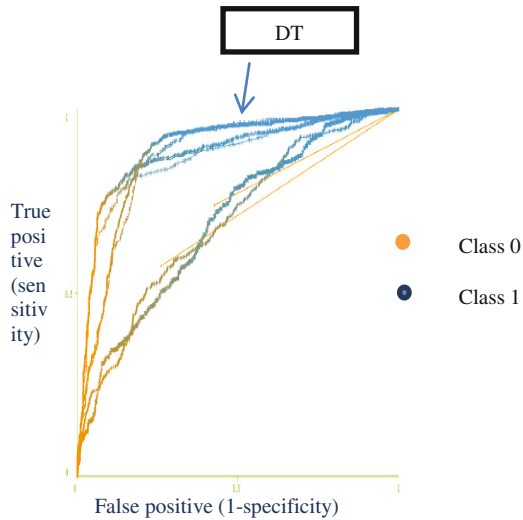


Fig. 9. ROC curve for progression class processed data

Table 9. Class progression result with unprocessed dataset (base model) and proposed dataset

	Base model				Proposed model			
	DT	MLP	NB	SVM	DT	MLP	NB	SVM
Accuracy (%)	88.0	83.0	86.9	88.4	81.7	77.9	62.7	66.6
Specificity	1.0	0.9	0.9	0.9	0.7	0.7	0.7	0.7
Sensitivity	0.0	0.08	0.0	0.0	0.8	0.8	0.5	0.5

7 Ranking of Attributes: Result and Analysis

Table 10 shows the ranks of the various attributes. Ranks are calculated by the amount of the information gain of the attributes. Treatment is the attribute which has highest rank in all the class except recurrence. In recurrence class the treatment has second rank. Cancer grade is an important variable in predicting whether the cancer would recur or not as its rank is 1 in recurrence class. Stage is a having significant contribution in prediction of the class death and the progression. Radiotherapy is a variable which has high contribution in predicting recurrence and metastasis. Surgery is an important factor that helps in predicting the recurrence and progression of the cancer. Age category is not a contributing factor for most of the classes, which mean age, has no significant impact in predicting all the classes except death. Cancer type is also not a contributing factor in most of the classes. This implies that whether the cancer is present in left, right or in both breasts, it may not affect the cancer recurrence, progression, death or metastasis.

Table 10. Ranking of each attribute in various classes

Class	Death	Metastasis	Recurrence	Progression
Attributes				
Age category	4	9	10	8
Cancer type	8	8	5	10
Grade	5	6	1	4
Stage	2	4	6	2
Estrogen-receptor (ER)	6	5	11	6
Progesterone receptor (PR)	7	11	8	7
Human epidermal growth factor receptor (Her2)	9	10	7	11
Status code	3	7	3	5
Treatment	1	1	2	1
Surgery	10	2	9	3
Radio therapy	11	3	4	9

8 Conclusion and Future Work

Various classification techniques have been applied on the raw data obtained from HCG hospital and the performance is studied. Similarly the same techniques algorithms are applied on the preprocessed and balanced data and performance is compared with result obtained from the raw data. Results significantly improved when the data is preprocessed and balanced. For balancing the data, sampling technique called SMOTE is applied which is the best suitable sampling technique for the dataset used in this work. MLP performs consistently well in most of the classes when sensitivity is considered. Ranking of the attribute is explored to understand the contribution of each attributes towards various classes various others over sampling technique can be

explored and checked whether there is increase in the performance. Association rule mining can be applied to extract relationship between various attributes.

References

1. Jothi, N., Wahidah, H.: Data mining in healthcare – a review. *Proc. Comput. Sci.* **72**, 306–313 (2015)
2. WHO Cancer - World Health Organization. <http://www.who.int/mediacentre/factsheets/fs297/en>
3. Cancer Statistics for the UK. <http://www.cancerresearchuk.org>
4. Khare, S., Gupta, D.: Association rule analysis in cardiovascular disease. In: *Second International Conference on Cognitive Computing and Information Processing (CCIP)*, SJCE, Mysuru, India, pp. 1–6. IEEE (2016)
5. Fan, Q., et al.: An application of apriori algorithm in SEER breast cancer data. In: *2010 International Conference on Artificial Intelligence and Computational Intelligence (AICI)*, vol. 3, pp. 114–116. IEEE (2010)
6. Gupta, D., Aggarwal, A., Khare, S.: A method to predict diagnostic codes for chronic diseases using machine learning techniques. In: *Fifth IEEE International Conference on Computing Communication and Automation (ICCA)*, pp. 281–287 (2016)
7. Dominic, V., Aggarwal, A., Gupta, D., Khare, S.: Investigation of chronic disease correlation using data mining techniques. In: *2nd International Conference on Recent Advances in Engineering and Computational Sciences (RAECS)*, pp. 1–6. University Institute of Engineering and Technology, Panjab University, Chandigarh (2015)
8. Dominic, V., Gupta, D., Khare, S.: Exploration of machine learning techniques for cardiovascular disease. *Appl. Med. Inf. Index Scopus* **36**(1), 23–32 (2015)
9. Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., Fotiadis, D.I.: Machine learning applications in cancer prognosis and prediction. In: *International Conference Science Direct*, pp. 8–17 (2014)
10. Sharma, N., Om, H.: Data mining models for predicting oral cancer survivability. *Netw. Model. Anal. Health Inf. Bioinform.* **2**(4), 285–295 (2013)
11. Yang, H., Chen, Y.P.P.: Data mining in lung cancer pathologic staging diagnosis: correlation between clinical and pathology information. *Expert Syst. Appl.* **42**(15), 6168–6176 (2015)
12. Abreu, P.H., et al.: Predicting breast cancer recurrence using machine learning techniques: a systematic review. *ACM Comput. Surv. (CSUR)* **49**(3), 52 (2016)
13. Kim, W., et al.: Development of novel breast cancer recurrence prediction model using support vector machine. *J. Breast Cancer* **15**(2), 230–238 (2012)
14. Ahmad, L.G., Eshlaghy, A.T., Poorebrahimi, A., Ebrahimi, M., Razavi, A.R.: Using three machine learning techniques for predicting breast cancer recurrence. *J. Health Med. Inf.* **4** (124), 3 (2013)
15. Park, K., et al.: Robust predictive model for evaluating breast cancer survivability. *Eng. Appl. Artif. Intell.* **26**(9), 2194–2205 (2013)
16. Sain, H., Purnami, S.W.: Combine sampling support vector machine for imbalanced data classification. *Procedia Comput. Sci.* **72**, 59–66 (2015)
17. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002)
18. Roozbahani, Z., Katanforoush, A.: Classification of gene expression data using multiple ranker evaluators and neural network. In: *CICIS*, pp. 29–31 (2012)

19. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco (1993)
20. Pal, S.K., Mitra, S.: Multilayer perceptron, fuzzy sets, and classification. *IEEE Trans. Neural Netw.* **3**(5), 683–697 (1992)
21. John, G.H., Langley, P.: Estimating continuous distributions in Bayesian classifiers. In: *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 338–345. Morgan Kaufmann Publishers Inc. (1995)
22. Platt, J.C.: 12 fast training of support vector machines using sequential minimal optimization. *Adv. Kernel Methods* **1**, 185–208 (1999)
23. Hanley, J.A., McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**(1), 29–36 (1982)

Performance Assessment Framework for Computational Models of Visual Attention

Bharathi Murugaraj^(✉) and J. Amudha

Department of Computer Science and Engineering, Amrita University, Bengaluru, India
bharathi02.subra@gmail.com

Abstract. This paper presents performance framework for computational model of visual attention, a software package, written using python scripting language, developed for the real-time comparison of computational model with human fixations. The performance framework was developed for real-time processing of eye trackers recorded data, analyzing them to generate fixation map, and comparing the fixation map to a saliency model got by running a configured computational model either in bottom-up or top-down mode. The framework is designed such that added modules can be extended for various experiment processing as required by the researcher. The framework encompasses the main connection to eye tracker to collect the raw data that will have observers eye coordinates and duration, it has analysis model to analyze the model and providing methods of visualization like fixation, heatmap and scanpath, it also has a computational model that predicts the fixation on the given image stimulus, finally the platform compares the fixation and saliency map to assess the accuracy of the prediction. All the functions of the framework can be controlled by using the graphical user interfaces.

Keywords: Eye tracker · Eye movement analysis · Fixation · Heatmap · Computational model · Saliency map · Python · Visual attention

1 Introduction

Visual attention mechanisms are developed that is derived from the study of human visual systems. Its process enables machine vision systems to select the most relevant regions from a scene Eye tracking is used for usability and psychology testing and used popularly in researches in visual system, cognitive process, and human-computer interaction. In psychophysiological research, eye-tracking methodology is used to get reaction parameters from eye movement data that used to analyze cognitive processes underlying visual behavior. Researcher's use eye movement data to study cognitive influence in learning, memory and attention. To perform eye tracking experiments many commercial eye-trackers e.g. EyeTribe [6], SMI [9], and open-source solutions e.g. Gaze Tracker [10] are available in market. They provide strong algorithms for gaze tracking, analysis and visualizations. Such tools provide valuable insights into the recorded gaze behavior. On the other hand some researchers have focused on the model-based analysis tools e.g. GazeAlyze [2], The

Psychtoolbox for Matlab [11], PsychoPy [12], for gaze reading in experimental studies in vision research. These tools come up with various detecting events in the gaze data, such as algorithms for blink detection, fixation detection etc.

Modeling visual saliency has attracted much interest recently and there are now several frameworks and computational approaches available. Some are inspired by cognitive findings, some are purely computational, and others are in between. Using wide variety of approaches many computational models of visual attention have been developed to see in a free-viewing condition how to predict where people look in images. The objective of these models is to improve artificial vision systems by computing, a numerical value of the likelihood of attending to, or the saliency of every location in an image. The performance of a model is measured by how well it predicts where people look in images in a free-viewing condition. So far, researchers validate prediction of attention models by direct comparison between eye movements recorded from humans watching the stimuli and model output. The fixations from the humans and the saliency map from the models is compared to assess the performance of the models.

To combine the eye tracking device, analysis of gaze data, saliency model and assessment of the model for its performance we developed a performance assessment framework. The framework provides 4 module: experimental module that helps to control experiment with the eye tracking device and record the eye movement data called gaze data of the observer, analysis module which generates a fixation map by analyzing the gaze data, saliency module generates a saliency map from a computational model (bottom-up/top-down), and performance comparator that compares the fixation map and saliency map using 3 popular metrics Normalized Scanpath Saliency (NSS), Pearson's Correlation Coefficient (CC) and Similarity.

2 Motivation

Eye tracking is process of finding out where the user is looking at for the given stimulus, by usage of eye tracking device. All eye tracking system function with a common principle, identifying the same eye features across the multiple images. And the results are correlated to a particular eye. Salient objects generally appear visually different from the other displayed objects in the scene. Eye tracking devices records eye coordinates and duration about a position of an eye within an eye's image as registered by a camera. This raw data is translated into a gaze point. For the computational tasks of recorded data analysis: fixation detection, heatmap and scanpath are considered. The most fundamental translation from raw data is the fixation detection, which is true for one eye tracker and for a specific dataset.

Generally a computational model handles several features and then computes them in parallel. The resulting value is fused in a representation called saliency map. The saliency map is visualized as a grey-scale image. In this map more the brightness of a pixel, it is the most salient region. Where human look in an image is based on two factors; a bottom-up and top-down approach. Bottom-up models are task-free and is stimulus-driven. Thus they do not require to learn, train or tune to open-ended task. They can be used for prediction on any image dataset, the output of it can be verified against

experimental data collected from humans. Top-down models are descriptive, task-specific and can be implemented computationally. But with lot of work going on this area the difference between bottom-up and top-down models is diminishing and there are algorithms of computational models that uses both the approaches for prediction. The paper [5] explores if existing object information can be used to for the next object recognition, it attempts to combine top-down and bottom-up model. Saliency maps produced by different algorithms are often evaluated by comparing output fixated image locations appearing in human tracking data. The inherent ambiguity in how saliency and ground truth are represented leads to different choices of metrics for reporting performance. Here, the performance comparator uses saliency metrics, that is, functions that take two inputs representing eye fixations and predictions and then output a number assessing the similarity or dissimilarity between them.

We presents a software performance assessment framework which is developed using python tool for comparing human fixations and the saliency map from the computational model for a given image stimulus. This will help in choosing appropriate models for the given application in hand. It can use various computational algorithm according to choice of the researcher and is an open-source software. Our intent was to developing an open-source, in-line python application that allows the complete management of entire processes of collecting data from eye tracker, post analyzing the collected data, predicting the salient region using either bottom-up or top-down models and finally compare them, that is, the fixation map and the saliency map.

3 Literature Survey

Even though, there are some well performing commercial eye trackers available, they have two disadvantages compared to the open-source solutions: they are either available at very high costs and thus becomes unaffordable for many academic research or clinical studies. They also provide a tightly-coupled environment that it is difficult to customize the way they need it for their application. Mostly all the solutions provide are black-box ensuring that there no access to image processing routines or modules. Many researchers explore different computational model to measure the accuracy of their prediction. The experiments conducted on 39 observers free-viewing 300 images and compare 10 popular recent modules [3]. It explores which models perform poorly and which ones are better. The performance of saliency models is measured using three different metrics ROC, similarity and EMD. Lot of paper explore on different evaluation metrics. Properties of the inputs affect metrics differently: how the ground truth is represented; whether the prediction includes dataset bias; whether the inputs are probabilistic; whether spatial deviation exist between the prediction and ground truth. Knowing how these properties affect metrics, and which properties are most important for a given application can help with metric selection for saliency model evaluation [4]. In this paper we use Normalized Scanpath Saliency (NSS), Pearson's Correlation Coefficient (CC) and similarity for fairest comparison of fixation map and saliency map. An open-source integrated framework like Visual Search Examination Tool (Vishnoo) [1] combines

configurable search tasks with gaze tracking capabilities, thus enabling the analysis of both, visual field and visual attention. This offers easily adaptable stimulus presentation, eye-tracking and evaluation of the visual scan path combined in a single platform.

4 System Architecture of the Framework

The frameworks block diagram is shown in Fig. 1.

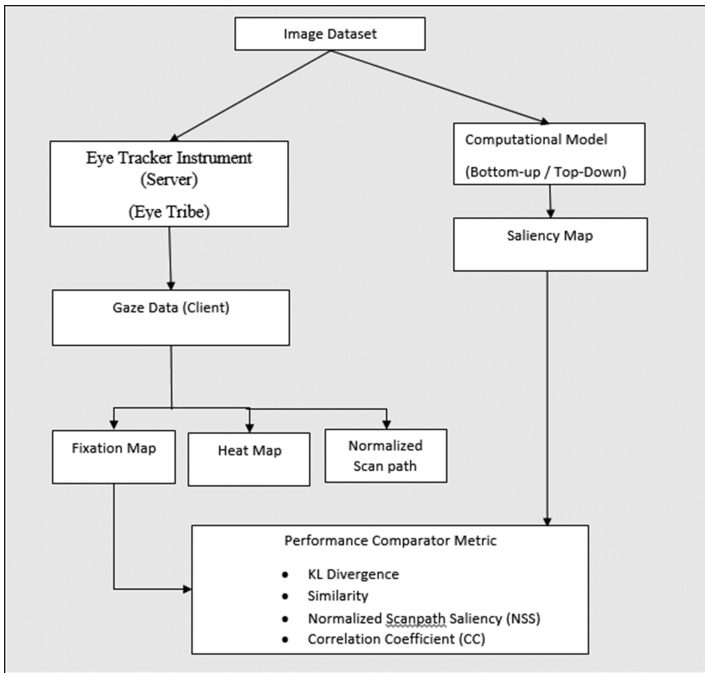


Fig. 1. Flow chart of performance framework

Researchers can use this framework tool with no specific programming skill, as it is graphically driven and all system parameters and all functionalities are controllable from graphical user interface (GUI). All system parameters of specific experiments are saved in a configuration script. This allows comfortable handling of the analysis of different experiments. The performance framework is entirely developed using python scripting language as it is open-source and can be easily portable. It runs on pretty much everything. All functions are written in python and are separate from the GUI components. The GUI is developed using Java programming language and is used explicitly for adding, updating and deleting configuration parameters. The performance framework consists of Experimental module, Analysis module, Saliency module and Comparator module.

Experimental Module

Eye tribe [11], an eye tracking device is used in the framework. By using the person’s face and eyes the device can calculate the exact location as to where the observers is looking at. The gaze coordinates are represented by a pair of (x, y) coordinates that is taken from the screen coordinate system, and it is calculated with respect to the screen the person is observing. Before using the eye tracker device the user need to do calibration process. Because the eye characteristics of individual is different, and this need to be modeled by the eye tracking software in order to estimate gaze accurately. Pygaze [8] an open-source python package is used in the framework. Pygaze acts as a wrapper around several existing eye tracking packages. It is used to create complicated experiments. Pygaze module connects to the eye tribe and records eye movement information like the eye coordination and duration into a tab separated file.

Analysis Module

Further the gaze data is analyzed by the Analyzer module which provides valuable insights into the recorded gaze behavior. In eye tracking data analysis fixation detection is considered. The velocity at each gaze point is calculated when the eye gaze travels from the previous gaze point to the current one. If calculated velocity is smaller than the threshold velocity then that gaze point is tagged as fixation. After completing tagging all the gaze points, consecutive fixation points are segregated into fixation groups <x,y,t,d> where x, y are center coordinates of the fixation group, t is the timestamp of the initial fixation point, and d is the duration. A fixation group is ignored for which the duration threshold is greater than the duration of a group.

Saliency Module

The first computational model of visual attention using the bottom-up approach was given by Koch and Ullman [13]. Here the visual features used were color, intensity or orientation. These feature maps are weighed and summed up to a saliency map.

Performance Module

If fixation map (FMap) and saliency map (SMap) are passed two inputs to a metric function, then the following metric are applied on them for comparison.

Normalized Scanpath Saliency (NSS): Measuring the normalized saliency at the region of fixations

$$NSS(SMap, FMap) = \frac{1}{N} \sum_i \overline{SMap}_i \times FMap_i$$

where $N = \sum_i FMap$ and $\overline{SMap} = \frac{SMap - \mu(SMap)}{\sigma(SMap)}$

Here i is the i^{th} pixel, and N is the total fixated pixels in fixation map. A positive NSS score indicates a good correspondence for a fixation located on the model predicted saliency map and the scanpath of the observer’s.

Similarity (SIM): Measuring the intersection between distributions

$$SIM(SMap, FMap) = \sum_i \min(SMap_i, FMap_i)$$

where $\sum SMap = \sum FMap = 1$

SIM measures the amount of similarity that exists between two distributions. A SIM of 1 means the FMap distribution and SMap distributions are same and SIM of 0 means there is no overlap.

Pearson's Correlation Coefficient (CC): Evaluating the linear relationship between distributions.

$$CC(SMap, FMap) = \frac{\sigma(SMap, FMap)}{\sigma(SMap) \times \sigma(FMap)}$$

where $\sigma(SMap, FMap)$ is the covariance of saliency map and fixation map. CC can range between -1 to 1 . CC is 1 means a perfect correlation, whereas -1 also means perfect correlation but in opposite directions.

5 Implementation

The GUI panel looks like in Fig. 2. All the system parameters are loaded from the configuration script. This can be changed or updated and rewritten to the same script or to a new configuration file.

To conduct the experiment, eye tracking system Eye tribe is used. Pygaze connects to the eye tribe server. It is prerequisite that the eye tribe need to be calibrated separately. Currently it is outside the scope the performance framework. Pygaze starts loading the dataset, that is, the collection of images in the gap of 30 s each. There are 48 images in the traffic sign dataset. They are in PNG format of size 360×270 , each of the image representing a traffic scene. In the dataset the image is categorized into 3 classes of different traffic sign template, pedestrian crossing, intersection and compulsory for bikes. The traffic sign dataset of 48 images was used to conduct the experiment to collect the eye movements from the participants who were allowed to free view each image for 2 s. Once this data is collected, the ground truth eye fixations from different participants is formatted as a column data and written to a tab separated file. The "rawx" and "rawy" data which represents the gaze coordinates. The Gaze coordinates are the point on screen that the user is currently looking. Gaze coordinates are defined as pixels in a top-left oriented 2D coordinate system and are available in both raw and smoothed forms. There are other information also in the tab separated file like "Tracking state", "Fixation" and "pupil coordinates". But for generating fixation Gaze coordinates values are used. The file content of tab separated file from the pygaze is show below in Fig. 3.

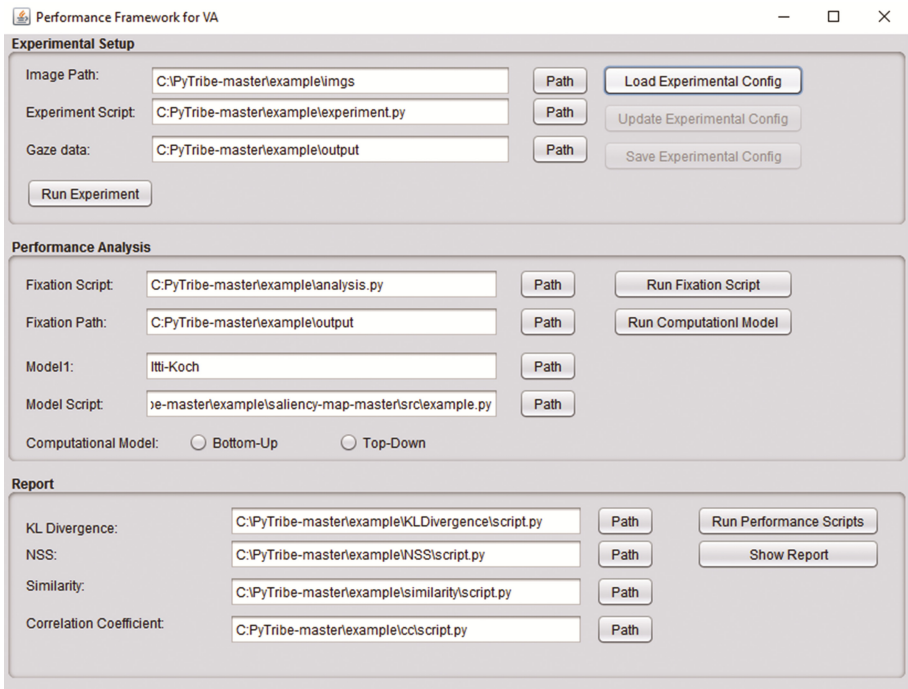


Fig. 2. The GUI of the performance framework

timestamp	time	fix	state	rawx	rawy	avgx	avgy	peize	Lrawx	Lrawy	Lavgx	Lavgy	lpeize	lpeixlx	lpeixly	Rrawx	Rrawy	Ravgx	Ravgy	
MSG	2017-05-09 14:58:09.916	747212648																		
MSG	2017-05-09 14:58:09.918	747212648																		
MSG	2017-05-09 14:58:09.918	747212648																		
2017-05-09 14:58:09.897	747212747	True	7	209.9456	14.5481	211.7244					70.1555	22.69535	249.498	142.5981			216.5096	130.2		
2017-05-09 14:58:09.931	747212747	True	7	215.4047	79.8809	212.2451					71.2595	23.17175	212.143	129.6109			216.1592	130.3		
2017-05-09 14:58:09.966	747212782	True	7	210.3596	98.6481	212.0662					74.3523	23.11075	259.9395	205.3217			219.9981			
2017-05-09 14:58:09.997	747212813	True	7	183.6211	61.1352	209.0484					72.879	22.25645	205.7461	149.6472			218.964	138.2		
2017-05-09 14:58:10.031	747212847	True	7	222.7424	46.0333	210.3749					70.1433	22.48885	220.7264	168.5596			219.1959			
2017-05-09 14:58:10.064	747212880	True	7	215.7993	113.4473	210.8767					74.0154	23.18365	227.6954	128.4629			219.9			
2017-05-09 14:58:10.097	747212913	True	7	249.9599	90.5681	214.2818					75.4948	23.31165	293.7928	154.3878			225.3577			
2017-05-09 14:58:10.131	747212947	False	7	376.5714	132.1481	376.5714					132.1481		21.754	400.7025			249.2847	400.7		
2017-05-09 14:58:10.164	747212980	False	7	591.0018	316.3358	591.0018					316.3358		22.60425	585.7699			409.3979			
2017-05-09 14:58:10.197	747213013	False	7	562.2266	277.2384	576.5988					296.7621		21.9854	529.3099			325.415	557.5139		
2017-05-09 14:58:10.231	747213047	False	7	580.295	235.9774	577.8098					276.3813		21.2724	492.4186			312.9564	535.6752		
2017-05-09 14:58:10.264	747213080	False	7	591.3182	268.1657	581.2036					274.2069		21.80985	526.3854			349.1411			
2017-05-09 14:58:10.297	747213113	False	7	590.1255	414.9236	583.0254					302.6876		22.8327	573.4144			465.2136	541.2		
2017-05-09 14:58:10.331	747213147	False	7	587.5638	292.9449	587.5638					292.9449		22.0671	612.0131			406.3149	612.0		
2017-05-09 14:58:10.364	747213180	False	7	739.0468	250.6187	739.0468					250.6187		21.69825	709.7431			368.4165			
2017-05-09 14:58:10.397	747213213	False	7	737.7656	338.6636	739.4948					294.6973		22.05075	685.9474			429.5345			
2017-05-09 14:58:10.430	747213246	False	7	706.1771	294.1787	727.6156					294.5984		21.6725	671.3132			346.9656	688.9		
2017-05-09 14:58:10.464	747213280	False	7	744.5189	258.6804	731.8373					285.5937		21.1437	692.8396			311.7455	689.8		
2017-05-09 14:58:10.497	747213313	False	7	711.416	243.3097	727.683					276.9857		20.7477	683.1076			688.4641	340.7		
2017-05-09 14:58:10.531	747213347	True	7	729.6401	180.1928	727.9761					260.4021		20.6393	749.5291			172.585	749.5291		

Fig. 3. Tab separated file generated from the pygze module.

Using the gaze coordinate and the duration value from the file is read and processed to generate fixation location and fixation map as shown in Fig. 4.

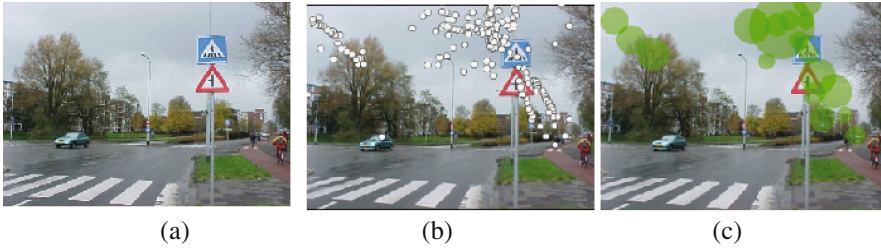


Fig. 4. (a) Traffic sign sample image from the dataset with its (b) fixation location and (c) fixation map

In implementation, free viewing task is considered since this makes it easier to use saliency models with fewest assumptions of the parameters. Two computational models of visual attention was used. Firstly basic Itti and Koch model based was used. The fixation location and fixation map as predicted by the model is displayed in Fig. 5.

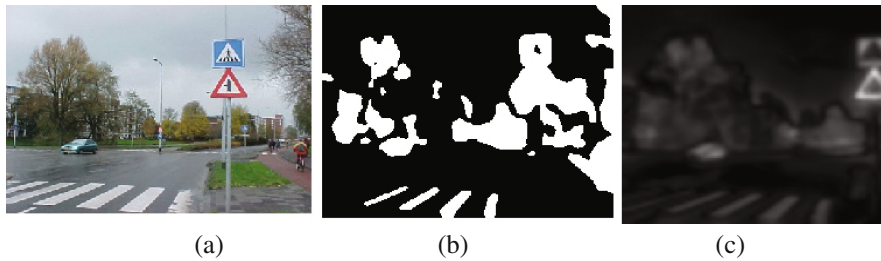


Fig. 5. (a) Traffic sign sample image from the dataset with its (b) binary image of salient regions and (c) its model saliency map using Itti-Koch

Graph Based Visual Saliency (GBVS) [16] was the second model used against the same image dataset. The saliency map is shown in Fig. 6.



Fig. 6. The GBVS model output (a) original image (b) saliency map

By overlaying the resulting salient regions from Itti-koch and GBVS, the original image is as shown in Fig. 7.

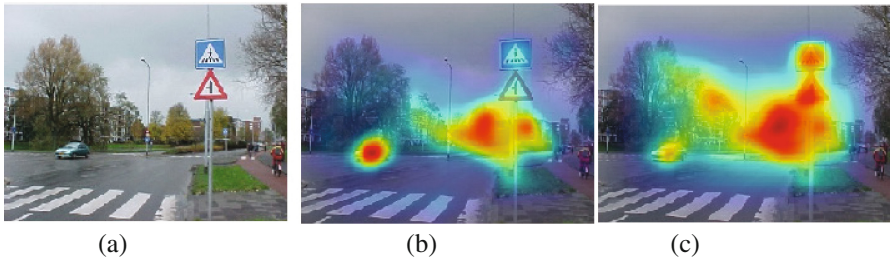


Fig. 7. Salient region overlaid on the original image (a) original image (b) Itti-koch (c) GBVS

The saliency model is assessed for validation using three different metric by the perfComparator: Similarity, Normalized Scanpath Saliency (NSS) and Correlation Coefficient (CC) as displayed in Fig. 8 above. After running the pygaze for the eye tracker experiment, and pygaze analyzer to generate fixation map, a comparison report is generated. This will compare between the fixation map and saliency map as show in the performance framework UI. The metric result is also displayed in the “Performance Metrics” UI table in Fig. 9. This gives a direct picture to the researcher as to how well the model is able to predict in compare to the ground truth.



Fig. 8. Performance measure of Itti-koch and GBVS model

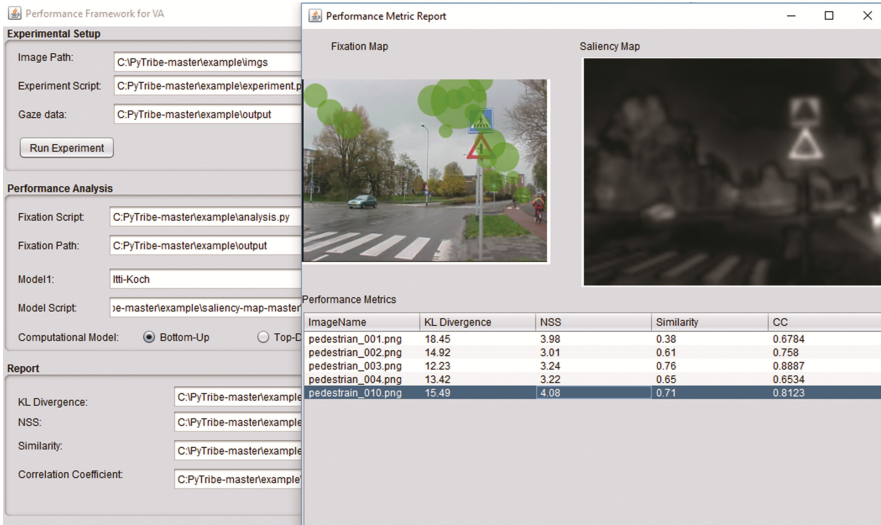


Fig. 9. The performance metric display comparing the fixation map and saliency map

6 Experiments

From the conducted experimental analysis we see that GBVS model performs better than the Itti-kotch. The Linear Correlation Coefficient (CC) values for the images are closer to 1, that is to say, there is a linear relationship exists between fixation map and saliency map. The similarity score is also approximating to 1, which tells us that the two maps distributions are same. This might be due to the factor that GBVS model also considers is center surround along with color, intensity and orientation in computing saliency map.

7 Conclusion

The performance framework provides a new platform for the researchers to experiment with wide range of performance assessment of computational models, and as well by collecting the human fixations from the eye tracking system module available in the framework. In Vishnoo top-down model used for specific task programming is assessed against scanpath i.e., focus of study is more on gaze data analysis, while the performance framework allows to compare the saliency map and fixation map using 3 popular metrics. Since all components required to test the performance of computational models is available as a combined solution in a single framework, which makes it fastest and flexible solution for researchers. Since the framework can be easily configurable, it becomes an attractive tool to many scientific research.

References

1. Tafaj, E., Kubler, T.C., Peter, J., Rosenstiel, W., Bogdan, M.: Vishnoo-an open-source software for vision research. In: 2011 24th International Symposium on Computer-Based Medical Systems (CBMS) (2011)
2. Berger, C., Winkeles, M., Lischke, A., Hoppner, J.: GazeAlyze: a MATLAB toolbox for the analysis of eye movement data. *Behav. Res. Methods* **44**, 404–419 (2012)
3. Judd, T., Durand, F., Torralba, A.: A benchmark of computational models of saliency to predict human fixations. MIT Technical report (2012)
4. Riche, N., Duvinage, M., Mancas, M., Gosseling, B., Dutoit, T.: Saliency and human fixations: state-of-the-art and study of comparison metrics. In: 2013 IEEE International Conference on Computer Vision (2013)
5. Amudha, J.: Performance evaluation of bottom-up and top-down approaches in computational visual attention system, Coimbatore (2012)
6. THEEYETRIBE: <http://www.theeyetribe.com>
7. Radha, D., Amudha, J., Jyotsna, C.: Study of measuring dissimilarity between nodes to optimize the saliency map. *Int. J. Comput. Technol. Appl.* **5**(3), 993–1000 (2014)
8. Dalmaijer, E.S., Mathot, S., Van der Stigchel, S.: PyGaze: an open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments. *Behav. Res. Methods* **46**, 913–921 (2014)
9. SensoMotoric Instrument GmbH (2011). <http://www.smivision.com>
10. <http://gazegroup.org/downloads/23-gazetracker/>
11. Brainard, D.H.: The psychophysics toolbox. *Spat. Vis.* **10**(4), 433–436 (1997)
12. Peirce, J.W.: PsychoPy-psychophysics software in Python. *J. Neurosci. Methods* **162**(1–2), 8–13 (2007)
13. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. *Hum. Neurobiol.* **4**, 219–227 (1985)
14. Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., Torralba, A. MIT saliency benchmark. <http://saliency.mit.edu/>
15. Amudha, J., Radha, D., Deepa, A.S.: Comparative study of visual attention models with human eye gaze in remote sensing images. In: Proceedings of the Third International Symposium on Women in Computing and Informatics (2015)
16. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Proceedings of Advances in Neural Information Processing Systems (NIPS), vol. 19, pp. 545–552 (2007)

Structural Matching of Control Points Using V-D-L-A Approach for MLS Based Registration of Brain MRI/CT Images and Image Graph Construction Using Minimum Radial Distance

Hema P. Menon^(✉) and A.S. Nitheesh

Department of Computer Science and Engineering, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, Amrita University, Coimbatore, India

p_hema@cb.amrita.edu, nitheeshas91@gmail.com

Abstract. Image registration using Moving Least Squares (MLS) is a point based method and requires the selection of control points from the source and target images. The selected points need not necessarily be the corresponding points. These points have to be matched to identify the corresponding points between the two images. For this, a new structural method for control point matching method is proposed. For structural matching the control points are represented using a graph structure and the structural properties like the degree of the vertex, length of edges and the angle between edges are used for finding the corresponding points in the source image and the target image. This method is found to be efficient for both mono-modal and multi-modal image registrations, as the topological property represented by the control points are exploited instead of the traditional intensity feature. The accuracy of the registration is computed using the standard Target Registration Error (TRE) Measure and compared with the registration using Thin Plate Splines (TPS). This work also proposes a new approach for constructing image graphs called as Minimum-Radial Distance (MRD) method.

Keywords: Image registration · Feature points · Moving least squares (MLS) · Structural matching · Target registration error (TRE) · Image graphs

1 Introduction

Image registration, a processing requirement for combining images, i.e., image fusion, is used extensively in the medical field for monitoring disease progression, image guided surgery, studying brain shift after surgery and for developing medical atlases. The process of image registration involves two images, namely, the source S and the target T images, and finding the best deformation field M to align these two images i.e., $S = M(T)$. After the registration problem is solved, for each pixel in the Source image a corresponding pixel in the target image is got. There is a scarcity of literature on efficient methods for selection and matching of control points in case of medical images

which are non-rigid or semi-rigid in nature. The number of control points taking part in the registration also influences the computation time taken for the registration process. Therefore, it is deemed worth exploring various methods including exploiting structural information of images for feature detection and matching that are not commonly adopted as of now. Of late the MLS deformation technique has been applied for registration of medical images. Navid Samavati [1] studied the issue of deformable liver image registration between Magnetic Resonance (MR) and Ultrasound (US) images of the liver. Author in this study has used Finite Element Modeling and Moving Least Squares for registration of images. The MLS deformation process has been detailed by Schaefer et al. [2] and is modified on the as-rigid-as-possible technique described by Igarashi and Hughes [3]. Numerous researchers have used the TSP transformations for image registration. The TPS algorithm was used to obtain an elastic transformation to map the source image to the target image by Rohr et al. [4, 6] and Cum et al. [8]. Construction of medical atlases for study purposes has been an area of focus in medical image processing and Park et al. [5] describes the use of TPS for such an atlas creation. Selection of control points for TPS by using the intensity as a control feature for consistent registration is discussed by Johnson and Christensen [7]. Registration can be done by comparing the intensity value, features present or the underlying image structure. There amount of literature available on intensity-based registration of images is enormous. Most researchers use mutual information (MI) as a measure for checking the image similarity. There are various methods for computing the mutual information between images [9–13]. Feature-based registration is applied mainly on task specific applications. Features are extracted from each image and then the correspondence is matched using any measure like mutual information or entropy [10, 12]. The common features that are extracted from an image are the Harris Corner point [14], Shi-Thomasi Corner points [15], the SURF points [16] and Edge points [17]. Rube et al. [18] has used SIFT features for extracting control points from aerial images. Structural approaches exploit the underlying structural information of images for feature matching. Structural information of an image can be represented using data structures like Laplacian Eigenmaps, Delaunay triangulation, Voronoi diagram and minimum spanning trees. Here the registration problem becomes a structure matching problem between two images [19–21]. A popular graph based registration approach deals with the concept of graph-cuts for registration. Another area of graph matching uses graph concepts for finding the matching [22, 23, 25]. The disadvantages of both the intensity and feature based methods can be reduced in structure based image registration. Since the structural information of a patch or an image is only dependent on the structures in the patch not the intensity values with which the structure is displayed. Registration is widely used to align image s in case of image reconstruction [25] and fusion [26].

2 Proposed Control Point (Feature Point) Matching Scheme Using Structural Similarity

The overall system design for the graph based structural control point matching scheme is shown in Fig. 1.

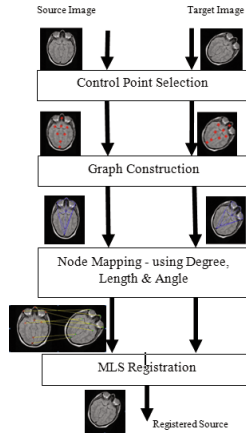


Fig. 1. Overall system design for graph based structural control point matching

2.1 Data Set Details

For the purpose of registration in this study we have used digital copies of MRI and CT images of human brain obtained as shown in Fig. 2.

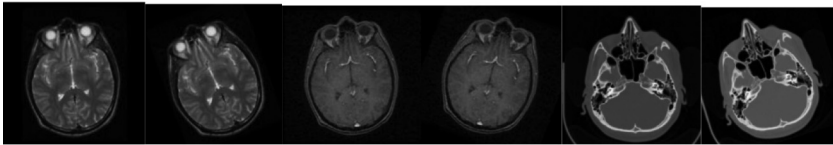


Fig. 2. Sample of MRI /CT brain image source-target pair

2.2 Structural Representation of Images Using Graphs

The image registration problem can be viewed as a graph matching problem once the image is represented as graphs. Hence graph construction becomes the basic step for performing registration. Image Graphs can be constructed in many ways. Mathematically, graphs can be represented by adjacency list, adjacency matrix, incidence matrix, Delaunay triangulation, minimum spanning trees, trees etc. The choice of graphs depends on the requirement of application for which it is used. In this work the Delaunay triangulation and Minimum-Radial Distance (MRD) method, which is a new approach proposed, are used for graph construction.

2.2.1 Image Graph Construction Using Delaunay Triangulation

For analysis and discussion the graph obtained. Figure 3 shows a sample output obtained for Delaunay graph constructed from the source and target image for selected points. The corresponding cell arrays showing the number of triangles in each image is given in Fig. 3. It can be noticed from the figure above that the number of

triangulations need not necessarily be the same. Delaunay eliminates very skinny triangles, thereby removing some of the selected points, which may pose a problem in case of medical images. A variant method has been proposed, named as ‘Minimum Radial Distance (MRD) Method’, which creates the same graph for an image on every execution and also all selected points are retained.

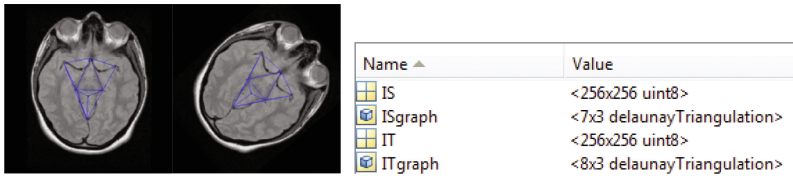


Fig. 3. Graph constructed using Delaunay triangulation on a sample source – target MRI image pair and adjacency and triangulation matrix obtained for graph.

2.2.2 Proposed Minimum Radial Distance Method for Graph Construction

Another simple way of constructing a graph from an image is to pass a circle with a predefined radius (obtained based on the image size) over the image, such that the center of the circle is on the current pixel ‘p’ to be processed. All pixels that fall within the circumference of the circle will be put into the neighborhood of ‘p’, and connected by a weighted edge. As the circle is passed on the image, all points in the image are thus interconnected. The edges are weighted by the distance and intensity measures. Thus, the image can now be represented using a weighted adjacency matrix. The steps of the algorithm are briefed below:

- (a) Considering each Control point as a node N in the image, visualize a circle of radius ‘r’ centered at the current point under consideration. The selection of the circle can be considered as a circular mask and the radius as the neighborhood size.
- (b) Find the nodes that lie inside the circle and form an edge between the initial node and all the nearby nodes.
- (c) Check if each node has at least 2 branches. This is to ensure that all points in the image form a part of the graph. If not, repeat the process with an increased radius for the circles, as the connectivity depends on this.
- (d) Repeat the step for every control point selected from both images.

For an image, let $G = \{V, E\}$ be the graph to be generated.

Where, $V = \{N_1, N_2, \dots, N_m\}$, $N_i \subseteq I$, and N_i be represented by the coordinates (a_i, b_i) for $i = 1, 2, \dots, m$.

$E = \{E_1, E_2, \dots, E_m\}$ and $E_k = \{all\ edges\ incident\ with\ N_k\}$

Initially $E_k = \emptyset$ for $k = 1, 2, \dots, m$.

- 1. Start with a node N_k to find the incidence edges to N_k .
- 2. Consider a circle C_r of radius r , defined by the following equation.

$$C_r : (x - a_k)^2 + (y - b_k)^2 = r^2$$

3. Find all nodes from V that fall within the circumference of C_r and establish an edge with N_k .

i. Compute the distance D_{ki}^2 between N_i and N_k

$$D_{ki}^2 : (a_i - a_k)^2 + (b_i - b_k)^2 \text{ for } i = 1, 2, \dots, m \text{ and } i \neq k, k = 1, 2, \dots, m$$

ii. If $D_{ki}^2 \leq r^2$ then we can have an edge e_{ki} between N_k and N_i

i.e., $e_{ki} = \{N_k, N_i\}$ and include this edge to $E_{k.}$, i.e., $E_k = E_k \cup \{e_{ki}\}$

iii. Repeat step 3 for $i = i + 1$

4. If $E_k \neq \emptyset$ then proceed with $k = k + 1$ and goto step 1

Else change the radius r such that radius $r_{new} > r$ and goto step 2.

5. Repeat steps 1 to 4 until $k = m$. The computed edges are weighted using the distance D_{ki} and the intensity difference L_{ki} between the intensity values of the nodes incident with the edge.

$$D_{ki} = \sqrt{(a_i - a_k)^2 + (b_i - b_k)^2}$$

$$L_{ki} = |I(N_k) - I(N_i)|$$

The graph obtained for the source-target MRI image pair and the adjacency matrix and degree matrix details is given in Fig. 4. Here $\{AdS, AdT\}$ are the adjacency matrices and $\{DS, DT\}$ the degree matrices of the source and target images respectively.

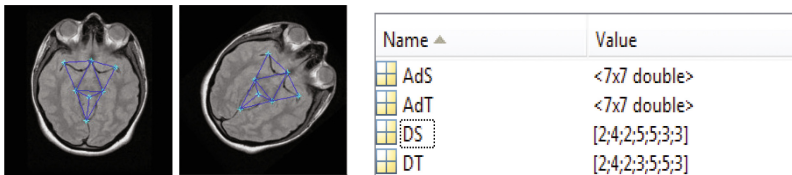


Fig. 4. Graphs created using minimum radial distance approach and adjacency and degree matrix obtained for graph

Let $G(V, E)$ be a graph with V as the set of vertices and E as the set of edges. Then, here we have two graphs $G_s(V_s, E_s)$ and $G_t(V_t, E_t)$ which are the graphs obtained for the source and target images. Where, V_s and V_t are the selected control points from source and target images and E_s and E_t are the Weighted edges between the neighbouring nodes of V_s and V_t . In the adjacency matrix A of a graph G , the non-diagonal entry a_{ij} is the number of edges from vertex i to vertex j , and the diagonal entry a_{ii} , is either once or twice the number of edges (loops) from vertex i to itself.

The Laplacian matrix of G , $L(G)$, is given by

$$x^T L(G)x = \sum_{(a,b) \in E} w(a,b)(x(a) - x(b))^2$$

$$L(G) = D(G) - A(G)$$

$$A(G) = \begin{cases} w(a,b) & \text{if } (a,b) \in E \\ 0 & \text{otherwise} \end{cases} \text{ which is the adjacency matrix}$$

$$D(G) = \text{diag}(d(a), a \in V) \text{ which is the diagonal matrix of vertex degrees of } G$$

These properties of graphs can be used to find the corresponding nodes in the two graphs. The graphs obtained are hence matched using the degree, edge weights length and angle between the edges.

3 Proposed Vertex-Degree-Length-Angle (V-D-L-A) Approach for Control Point Matching

The V-D-L-A method of control point matching is proposed mainly for registration involving multi-modal images and for mono-modal images where intensity variations between the source and target images are more. The graph matching is done by an iterative method which involves finding the degree of each vertex, length of edges of a graph and angle formed between the edges of the graph. The matched points thus obtained are then given to the Moving Least Squares (MLS) based registration technique.

3.1 V-D-L-A Control Point Matching Process

The steps involved in the V-D-L-A control point matching process are given below:

1. Create the Delaunay triangulation/Minimum Radial Distance based graph of the set of points from source as well as target image.
2. Formulate the adjacency and degree matrix of the connected list of points separately for both source and floating image.
3. For each image, find nodes with similar degree and enlist them along with the weights between these nodes.
4. Reduce the grouping in a one-to-one basis such that each node in source image corresponds to some specific node in the floating image by considering the pair which has the least error in length of branches between parent node and the neighbouring nodes among all the matching pairs.
5. Repeat step 4, now considering the angle between the branches for each and every node.
6. If considering only lengths or only angles, proceed to step 8. Else go to step 7.
7. Compare the matched list created by considering the lengths of branches as well as angles between branches and select the matched pairs which represent in both the matched lists. This is the final set of matched points.

8. Formulate a threshold by considering the median value of the angle or weight error set.
9. Discard all the nodes having an error greater than formulated threshold. The remaining set will be the matched set of points obtained using only lengths or angles.

4 Results and Discussions

The step by step results obtained for the structural method has been discussed here in this section.

1. After the control points are represented using graphs as shown in Fig. 5 formulate the adjacency and degree matrix of the connected list of points separately for both source and floating image. The edges connecting the nodes are weighted using Euclidian distance between the nodes that they connect. If the nodes are not connected then it is represented as '*inf*' and diagonal will be zero as it represents the node (i, i).

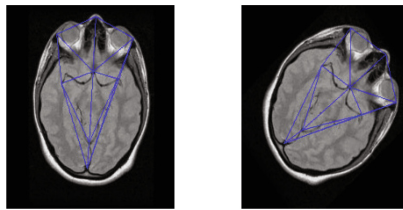


Fig. 5. Source and target image graph

All other cells in the matrix have a finite value depicting the distances. The weighted adjacency matrix thus obtained for the Graph in Fig. 5 are given below in Fig. 6(a) and (b) respectively.

Using values present in each cell the degree of each of the nodes is computed.

Degree (Node i) = no of cells in that row having finite values.

The degree matrix for the weighted adjacency matrices is shown in Fig. 7.

It can be observed from the degree matrices that one node in source image graph has many node with the same degree in the target image graph.

2. Find the nodes with the same degree to create the initial correspondence matrix as shown in Fig. 8.

Figure 8 shows that node 1 in Source image graph may correspond to any of the nodes namely the Nodes 1, 2, 4, 6 or 8 of Target image graph. By grouping the node based on the degree, the search space for the next level is reduced. The next step is to reduce the correspondence to one-to-one basis such that each node in source image corresponds to some specific node in the floating image.

	1	2	3	4	5	6	7	8
1	0	50.4356	Inf	63.8375	Inf	Inf	85.7846	112.0036
2	50.4356	0	38.3050	72.7815	73.0227	68.2540	89.9546	Inf
3	Inf	38.3050	0	Inf	55.9161	Inf	90.9618	120.9736
4	63.8375	72.7815	Inf	0	Inf	62.1866	Inf	Inf
5	Inf	73.0227	55.9161	Inf	0	61.6445	Inf	176.8825
6	Inf	68.2540	Inf	62.1866	61.6445	0	Inf	Inf
7	85.7846	89.9546	90.9618	Inf	Inf	Inf	0	31.3142
8	112.0036	Inf	120.9736	Inf	176.8825	Inf	31.3142	0

(a)

	1	2	3	4	5	6	7	8
1	0	90.8061	89.8205	85.9858	Inf	Inf	Inf	30.2047
2	90.8061	0	38.5383	Inf	Inf	55.7289	Inf	119.7731
3	89.8205	38.5383	0	49.2064	68.5862	72.9306	73.6163	Inf
4	85.9858	Inf	49.2064	0	Inf	Inf	64.1017	111.3335
5	Inf	Inf	68.5862	Inf	0	60.2587	63.8730	Inf
6	Inf	55.7289	72.9306	Inf	60.2587	0	Inf	175.4991
7	Inf	Inf	73.6163	64.1017	63.8730	Inf	0	Inf
8	30.2047	119.7731	Inf	111.3335	Inf	175.4991	Inf	0

(b)

Fig. 6. (a) Weighted adjacency matrix of the source image graph. (b) Weighted adjacency matrix of the target image graph

	1		
1	4	←	The node 1 in source graph has 5 nodes in target graph with the same degree
2	6	←	
3	4	←	
4	3	←	
5	4	←	
6	3	←	
7	4	←	
8	4	←	

Fig. 7. Degree matrix computed for the source and target image graphs.

- The step two has reduces the comparison space for each node. Now the comparison is done only between the nodes whose degrees have matched. Observing the weighted adjacency matrix given in Fig. 9 it can be seen that Node 1 of Source graph corresponds to Node 4 of Target graph. This correspondence can be achieved by considering the pair which has the least error in length of branches and angle between parent node and the neighbouring nodes among all the matching pairs. The Similarity matrix obtained using the length and angle similarity matrix is given in Fig. 10.
- Generate the separately matched pairs using length of branches and angle between branches. From the matched points obtained, check for matches that retains in both matched points set. Discard all pairs that are not present in both matched sets. The points thus matched is given in Fig. 11. Figure 12(a) shows the matched control

	1	2	
1	1		1
2	1		2
3	1		4
4	1		6
5	1		8
6	2		3
7	3		1
8	3		2
9	3		4
10	3		6
11	3		8
12	4		5
13	4		7
14	5		1
15	5		2
16	5		4
17	5		6
18	5		8

The 5 matches found for node 1

Fig. 8. Initial correspondence matrix based on degree similarity

	1	2	3	4	5	6	7	8
1	0	50.4356	Inf	63.8375	Inf	Inf	85.7846	112.0036
2	50.4356	0	38.3050	72.7815	73.0227	68.2540	89.9546	Inf
3	Inf	38.3050	0	Inf	55.9161	Inf	90.9618	120.9736
4	63.8375	72.7815	Inf	0	Inf	62.1866	Inf	Inf
5	Inf	73.0227	55.9161	Inf	0	61.6445	Inf	176.8825
6	Inf	68.2540	Inf	62.1866	61.6445	0	Inf	Inf
7	85.7846	89.9546	90.9618	Inf	Inf	Inf	0	31.3142
8	112.0036	Inf	120.9736	Inf	176.8825	Inf	31.3142	0

	1	2	3	4	5	6	7	8
1	0	90.8061	89.8205	85.9858	Inf	Inf	Inf	30.2047
2	90.8061	0	38.5383	Inf	Inf	55.7289	Inf	119.7731
3	89.8205	38.5383	0	49.2064	68.8662	73.0206	73.6163	Inf
4	85.9858	Inf	49.2064	0	Inf	Inf	64.1017	111.3335
5	Inf	Inf	68.8662	Inf	0	60.2587	63.8730	Inf
6	Inf	55.7289	72.9306	Inf	60.2587	0	Inf	175.4991
7	Inf	Inf	73.6163	64.1017	63.8730	Inf	0	Inf
8	30.2047	119.7731	Inf	111.3335	Inf	175.4991	Inf	0

Fig. 9. A sample node correspondence identified from the weighted adjacency matrix.

points on the image when considering length alone and Fig. 12(b) shows the matches obtained on considering the length and the angle and registered source is shown in Fig. 12(c).

This method was experimented CT-CT, MRI-MRI and MRI-CT. Sample results obtained using the V-D-L-A approach discussed applied on CT-CT image pair is shown in Figs. 13 and 14. It can be observed from the Fig. 14 that out of the many point, only a few points are matched. The matching obtained in case of multi modal images is shown in Fig. 15. Experiments were conducted on mono and multi modal images and the representative outputs that justify that MLS registration can be performed even with minimal number of correctly matched control points is shown in Figs. 16 and 17.

	1	2	3	4	5	6	7	8
1	0	20.2309	Inf	22.1484	Inf	Inf	0.2012	21.1974
2	0	5.2933	Inf	8.1086	Inf	Inf	5.0215	7.7696
3	0	1.2292	Inf	0.2643	Inf	Inf	0.2012	0.6701
4	0	5.2033	Inf	3.5788	Inf	Inf	12.8540	39.0730
5	0	20.2309	Inf	33.6327	Inf	Inf	25.5489	0.6701
6	1.2292	0	0.2333	0.1491	0.0921	0.3321	0.1341	Inf
7	Inf	8.1002	0	Inf	25.7113	Inf	0.1556	30.1675
8	Inf	0.2333	0	Inf	0.1872	Inf	0.1556	1.2005
9	Inf	10.9014	0	Inf	6.7097	Inf	4.9759	9.6402
10	Inf	17.4239	0	Inf	0.1872	Inf	18.0312	48.0430
11	Inf	8.1002	0	Inf	25.7113	Inf	20.3717	1.2005
12	0.0355	4.1954	Inf	0	Inf	1.6864	Inf	Inf
13	0.0355	0.8348	Inf	0	Inf	1.6864	Inf	Inf
14	Inf	12.8632	25.7113	Inf	0	24.3414	Inf	86.0763
15	Inf	17.2928	0.1872	Inf	0	5.0156	Inf	57.1092
16	Inf	8.9210	6.7097	Inf	0	2.4572	Inf	65.5490
17	Inf	0.0921	0.1872	Inf	0	1.3858	Inf	1.3834
18	Inf	38.3108	25.7113	Inf	0	31.4397	Inf	1.3834
19	Inf	0.3321	Inf	1.6864	1.3858	0	Inf	Inf
20	Inf	4.1523	Inf	1.6864	2.2285	0	Inf	Inf
21	0.2012	0.1341	0.1556	Inf	Inf	Inf	0	1.1094
22	5.0215	0.8515	0.1556	Inf	Inf	Inf	0	7.2241

	1	2	3	4	5
1	0.9045	0.9045	74.2975	74.2975	74.2975
2	2.9649	2.9649	76.3579	76.3579	76.3579
3	2.1059	1.2128	0.2045	1.0976	48.6130
4	0.5940	0.5940	74.2975	74.2975	74.2975
5	1.4664	1.4664	76.3579	76.3579	76.3579
6	0.0114	2.5131	2.2474	8.9530	0.5779
7	0.6380	0.4478	0.2263	9.6024	34.8231
8	1.1076	27.7526	2.8371	8.2255	42.3722
9	0.4581	15.7640	0.8160	8.8749	1.1444
10	3.0634	1.4868	0.2568	114.8625	114.8625
11	82.6304	64.5010	66.2446	23.4986	23.4986
12	50.4038	25.7100	24.6938	114.8625	114.8625
13	1.5291	0.3338	0.6823	23.4986	23.4986

Fig. 10. Length similarity matrix and angle similarity matrix

	1	2
1	1	4
2	2	5
3	3	6
4	4	7
5	5	8
6	6	1
7	7	2
8	8	3

Fig. 11. Matched points using length.

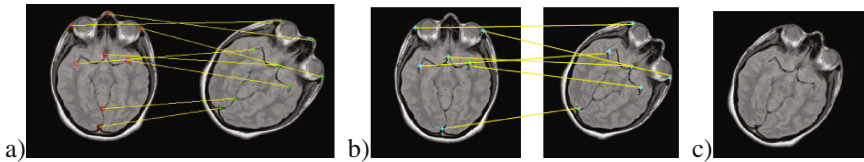


Fig. 12. Matched control points in the source and target image (a) using length alone (b) using length and angle (c) registered source

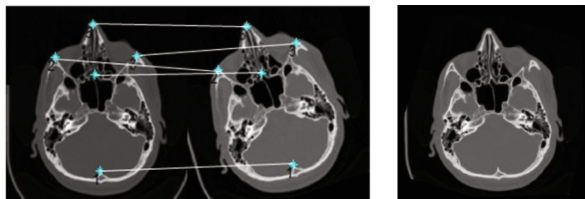


Fig. 13. Matched points for CT images with manually selected points and the registered image with TRE = 1.009

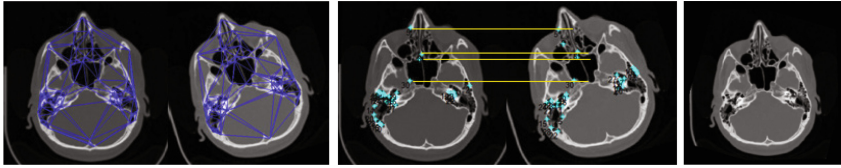


Fig. 14. Image Graphs with more number of points, matched points from graphs and the registered image with TRE = 1.098

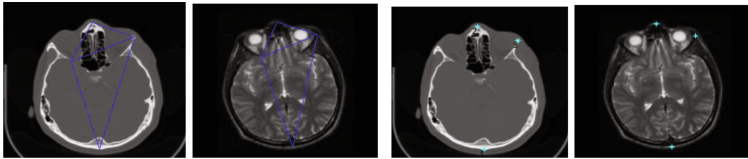


Fig. 15. Image graph created from few points in source CT and target MRI image and the matched points using V-D-L-A method.

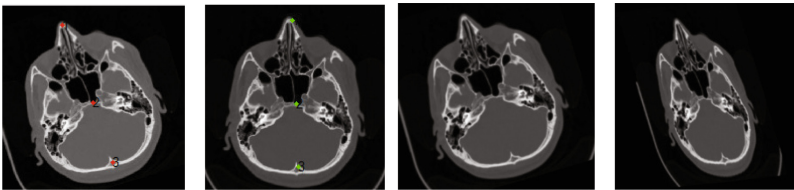


Fig. 16. CT image source – target pair with minimal matched points and the MLS transformed source image and the TPS transformed source image.

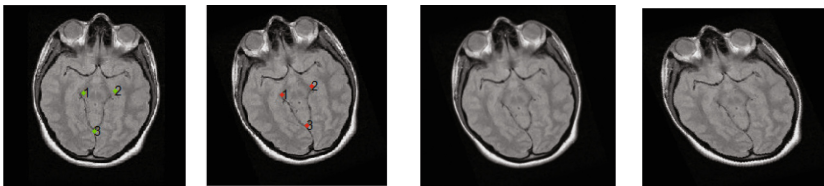


Fig. 17. MRI image source – target pair, with minimal matched points and the MLS and TPS transformed source image.

The results obtained are compared with the TPS registration method by computing the Target Registration Error (TRE) and tabulated in Table 1.

The ideal value of TRE is generally accepted as 1, in cases where lesser number of control points are selected and matched using V-D-L-A approach based MLS registration, the TRE is much lesser.

Table 1. Comparison of TRE for MLS and TPS registration.

No. of matched control points	Standard deviation of target registration error	
	MLS registration	TPS registration
3	1.6432	2.2689
4	1.3533	1.6567
5	1.4065	1.7250
6	1.4032	1.7156
8	1.2002	1.4132
10	1.2356	1.3786
21	1.1056	1.2442
40	1.0090	1.1104

5 Conclusion

In this work a structural method based on vertex, degree, length and angle properties of image graph was implemented and tested on MRI and CT images. This work also proposed a new approach for constructing image graphs called as Minimum-Radial Distance (MRD) method. The findings from this work are observed to be more applicable in medical image processing because in such application it is not practical to get large number of control points from the two images considered, owing to reasons like slice differentiation in images or missing data due to non-availability of exact image. Also, in medical image analysis, time taken for computation is crucial because of the inherent nature of medical related activities. Therefore, it is essential to have a technique working efficiently with lesser number of control points. And hence the correctness of the selected points is of great importance. Since the efficiency of the MLS registration depends on the corresponding control points, an automatic or manual control point selection followed by a structural control point matching was found to be more effective, and produced registered images with lesser TRE. The structural control point matching is suitable for both mono and multi modal images.

References

1. Samavati, N.: Deformable multi-modality image registration based on finite element modeling and moving least squares, MS thesis, Department of Electrical and Computer Engineering and the School of Graduate Studies, McMaster University (2009)
2. Schaefer, S., McPhail, T., Warren, J.: Image deformation using moving least squares. In: SIGGRAPH 06, ACM Transactions on Graphics, pp. 533–540 (2006)
3. Igarashi, T., Hughes, J.: As-rigid-as-possible shape manipulation. In: ACM Transactions on Graphics, vol. 24, pp. 1134–1141 (2005)
4. Rohr, K., Fornefett, M., Stiehl, H.S.: Spline-based elastic image registration: integration of landmark errors and orientation attributes. *Comput. Vis. Image Underst.* **90**(2), 153–168 (2003)

5. Park, H., Bland, P.H., Meyer, C.: Construction of an abdominal probabilistic atlas and its application in segmentation. *IEEE Trans. Med. Imaging* **22**(4), 483–492 (2003)
6. Auer, M., Regitnig, P., Holzapfel, G.A.: An automatic non-rigid registration for stained histological sections. *IEEE Trans. Image Process.* **14**(4), 475–486 (2005)
7. Johnson, H.J., Christensen, G.E.: Landmark and intensity based, consistent thin-plate spline image registration. In: *Proceedings of International Conference on Information Processing in Medical Imaging*, vol. 2082, pp. 329–343 (2001)
8. Crum, W., Hartkens, T., Hill, D.L.G.: Non-rigid image registration: theory and practice. *Br. J. Radiol.* **77**, 140–152 (2004)
9. Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., Suetens, P.: Multimodality image registration by maximization of mutual information. *IEEE Trans. Med. Imaging* **16**, 187–198 (1997)
10. Nyúl, L.G., Udupa, J.K., Saha, P.K.: Task-specific comparison of 3-D image registration methods. *Med. Imaging Image Process.* **4322**, 1588–1598 (2001)
11. Rodríguez-Carranza, C.E., Loew, M.H.: A weighted and deterministic entropy measure for image registration using mutual information. *Med. Imaging Image Process.* **3338**, 155–166 (1998)
12. Ioannides, A.A., Liu, L.C., Kwapien, J., Drozd, S., Streit, M.: Coupling of regional activations in a human brain during an object and face affect recognition task. *Hum. Brain Mapp.* **11**(2), 77–92 (2000)
13. Pompe, B., Blidh, P., Hoyer, D., Eiselt, M.: Using mutual information to measure coupling in the cardio respiratory system. *IEEE Eng. Med. Biol. Mag.* **17**, 32–39 (1998)
14. Harris, C.G., Stephens, M.J.: A combined corner and edge detector. In: *Proceedings of Fourth Alvey Vision Conference*, Manchester, pp. 147–151 (1988)
15. Shi, J., Tomasi, C.: Good features to track. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 593–600 (1994)
16. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: speeded up robust features. In: *Proceedings of 9th European Conference on Computer Vision*, Part 1, pp. 404–417 (2006)
17. Canny, J.: A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-8**(6), 679–698 (1986)
18. Rube, I.E., Sharkas, M., Salman, A., Salem, A.: Automatic selection of control points for remote sensing image registration based on multi-scale SIFT. In: *Proceedings of 2011 International Conference on Signal, Image Processing and Applications (SIA 2011)*. Chennai, India, 17–18 December 2011
19. Tian, J., Lee, N., Theodore, R., Smith, A., Laine, L.: A partial intensity invariant feature descriptor for multimodal retinal image registration. *IEEE Trans. Biomed. Eng.* **57**(5), 1707–1718 (2010)
20. Zheng, J., Tian, J., Deng, K., Dai, X., Min, X.U.: Salient feature region: a new method for retinal image registration. *IEEE Trans. Inf. Technol. Biomed.* **15**(2), 221–232 (2011)
21. Tsai, C., Li, C., Yang, G., Lin, K.: The edge-driven dual-bootstrap iterative closest point algorithm for registration of multimodal fluorescein angiogram sequence. *IEEE Trans. Med. Imaging* **29**(3), 636–649 (2010)
22. Holden, M., Hill, D.L.G., Denton, E.R.E., Jarosz, J.M., Cox, T.C.S., Rohlfing, T., Goodey, J., Hawkes, D.J.: Voxel similarity measures for 3-D serial MR brain image registration. *IEEE Trans. Med. Imaging* **19**, 94–102 (2000)
23. Freeborough, P.A., Fox, N.C.: Modelling brain deformations in Alzheimer disease by fluid registration of serial MR images. *J. Comput. Assist. Tomogr.* **22**(5), 838–843 (1998)
24. Talairach, J., Tournoux, P.: *Coplanar Stereotaxic Atlas of the Human Brain*, p. 1988. Thieme Medical, New York (1998)

25. Arathi, T., Parameswaran, L.: Image reconstruction from 2D stack of MRI/CT to 3D using shapelets. *Int. J. Eng. Technol. (IJET)* **6**, 2595–2603 (2014)
26. Bhavana, V., Krishnappa, H.K.: A survey on multi-Modality medical image fusion. In: *Proceedings of the 2016 IEEE International Conference on Wireless Communications, Signal Processing and Networking, WiSPNET 2016*, pp. 1326–1329 (2016)

An Interactive and Intelligent Tool for Circuit Component Recognition Through Virtual Reality

Shriram K. Vasudevan^(✉), S.N. Abhishek, N.K. Keerthana, Rajan Priyanka, A. Aravinth, and M. Divya

Department of Computer Science and Engineering, Amrita School of Engineering,
Amrita Vishwa Vidyapeetham, Amrita University, Coimbatore, India
kv_shriram@cb.amrita.edu

Abstract. The electronic products that support the needs in our daily life require a long drawn out process for its development and building. Every step from scratch i.e. requirement collection to integration and testing, involves a lot of time consuming tasks. Generally, a technically sound person with good knowledge in simulation is assigned to carry out these tasks, yet it fritters away time. The idea proposed here is an application for generating NetList for the circuit drawn through hand gesture, powered by virtual reality. The application gets input from the user's hand gesture using an additional hardware called leap motion sensor. This enables the users to draw any kind of circuit, of any size with any number of components. The completed circuit is captured and further processing is carried out. The circuit is isolated, recognized and segregation of the wires and components takes place in the consecutive phases. The next step involves identification of individual components and the aggregated NetList will be generated in the final step. And also, complex circuit which cannot fit into a limited space can be visualized in the virtual environment.

Keywords: Machine learning · Component recognition · Image processing · Electronic circuit · NetList · Virtual reality · Gesture

1 Introduction

Testing an electronic circuit is a tedious job. In the present times, it is done by virtually generating the circuit, with the help of programs and simulator. But again, it consumes lot of time and engineers find it difficult to write the perfect code for simulation. Most of the existing application requires images of the circuit to be fed in for stimulation, posing few constrains. Considering the above facts, we intend to create an application that can generate the code to the given circuit on the go i.e., the application focus on images that are drawn through hand gestures.

Virtual and augmented reality is a boon for the modern generation computing and has helped in developing a vast range of applications. Developers and Researchers have also found great potential and interest in image processing and are using it for simple applications to complex applications. This product is one such attempt where we make

machine to sense and read the input and process it as an image, understand the contents deeper and finally provide the expected output. Image processing and machine learning together are revolutionizing the digital world, whereas virtual/augment reality is making the impossible happen. Adding all the three, adds immense strength to the product.

Our application takes the image of an electronic circuit drawn using hand gestures and returns a netlist for the same. The process involves identification, recognition and grouping of the electronic component. To achieve accuracy various algorithms have been used. A special hardware, Leap motion sensor, is used to track the gestures and motion of the hand and records the same. Such a system has many advantages and will help save huge amount of timing in reconfirming the design for the electronic component. The novelty of our project is that we focus on images that are drawn through hand gestures, thus enabling the users to simulate on the go.

This project is a new attempt where we have tried to generate textual representation of the circuit by taking input in virtual reality. Usage of virtual reality is being the trend and it would add a lot of value to the developed application. If scaled to a higher level it has a great potential to ease the job testing of electrical circuits and stimulate it on the go. The concept of making the application more digitalized supports in reducing the usage of paper and pen. This makes the application versatile and can be used at any point of time. Moreover the application is independent and takes less time in generating the netlist compared to existing application [8, 9].

2 Literature Survey

Few research works have already been carried out in this area and there are quite a number of applications and platforms which takes the user's input and process it for stimulation and study of electronic systems. Some of the existing solutions and approaches that were found useful in gaining deeper insights in the proceedings to be carried out are listed below.

In the year 1997, Bailey et al., exhibited a system that scans printed circuits, recognize its components and generates a basic netlist. It is known to be the earliest of all the solution for this requirement and forms a base for further works. They have used the image processing technique for recognition of components. One of the restrictions is that, it deals only with circuits printed in books and not any other. In this modern era with lot of advancements in the area of image processing, this is not a challenging solution. Yet another pitfall cited is that it fails to recognize labels [1].

Feng and coauthors in the year 2008 proposed a technique for sketch understanding that automatically extract symbols from the continuous stream of strokes. This approach is based on a two dimensional dynamic programming technique (2D-DP) which allows to locate symbols even though they might be temporally overlapped with each other. The pro of this method is that unlike the previous works, it is not demand that symbols should be drawn step by step. From the conducted experiments accuracy is found to be more than 90%. Few cause of the recognition failure includes scribing errors, forgetting to draw a stroke of a symbol and a few more which they had mentioned to solve in future works [2].

The development of the application was also based on the methods discussed in “Hand-Drawn Circuit Recognition” by Ravi Palakodety and Vijay Shah, Massachusetts Institute of Technology. This work develops a tool for recognizing components and their values. It also finds the connectivity’s of components from a hand-drawn circuit to generate a SPICE netlist. It allows the user to draw the circuit on an 8×8 grid. The components have to be centred and sized to fill the grid blocks. Components like resistors, transistors, voltage sources, power supplies, capacitors, ground terminals can be used. The final output will be the replacement of the hand drawn circuit components with computer generated component pictures and the generated SPICE netlist. After studying the work, the component recognition was found to be 100% accurate and text recognition accuracy was up to 90%. The work analyses the hand drawn circuit and creates a netlist and the output is got in both textual and graphical format [3].

Sridar et al. followed an interesting approach in image processing along with some machine learning techniques and proposed an offline circuit recognition application. This combination and approach is the first among researches related to this problem. They have proposed a stimulation model where the foremost step is to draw the input circuit on a white sheet of paper which is then captured by a camera under good lighting conditions. De-noising is performed and the image is converted to grayscale using Otsu’s method. For netlist creation the components needs to be identified. The approach mentioned in this paper for identification of components is to find the patterns that are unique to each component. It is considered as one of the best as it gives accuracy of about 80% [4].

3 Problem Statement

The product proposed here is an application, which will allow the user to draw any kind of circuit, of any size with any number of components with free hands using gestures in a virtual environment. Once the components are identified and recognized, Netlist creation along with labels is obtained in the final stage which can be fed into stimulators.

4 Component Recognition

The components are given as images to the machine learning algorithm and the same is identified. This can be achieved by both supervised and unsupervised learning.

5 Supervised Classification

In supervised classification, is simple the act of concluding to the result from a labeled set of data. Here the range of outputs is previously determined before readying the model. Data set for each output class is fed in separately and data sets are properly labelled. This labelled data is called the training data. Once the labelled data set is ready and the algorithm is complete, the model can be deployed. When an input is received, it is classified based on the tagged training data. For circuit component recognition, the data set is the circuit components.

6 Unsupervised Classification

On the other hand unsupervised learning doesn't have any training data. Here, when the input comes the ML classifies the input into one of the available classes. The output is based on clusters. Clusters are formed by the algorithm using the inputs given. So, when a new input comes it will be checked with all available clusters and assigned to one of them. If it doesn't fall into any cluster, then the algorithm creates a new cluster and places the input in it. So, in component recognition at the end will be different classes (clusters). Each cluster will be one component and we should tag it later.

The best algorithms used for image classification are:

- K-Nearest Neighbors
- Support Vector Machines
- Decision Tree
- K-means
- Fuzzy C means

7 Algorithm selection

The results of these algorithms for various combinations of the circuit components are shared below in Table 1.

Table 1. Machine learning algorithms result for various combinations of RLC and voltage source circuit

Circuit	K-NN	SVM	Decision Tree	K-means	Fuzzy C means
Only Resistor (R)	R-jpg R	R-jpg R	R-jpg R	R-jpg Cluster 1	R-jpg Cluster 1
Only Inductor (L)	L-jpg L	L-jpg L	L-jpg L	L-jpg Cluster 1	L-jpg Cluster 1
Only Capacitor (C)	C.jpg C	C.jpg C	C.jpg C	C.jpg Cluster 1	C.jpg Cluster 1
Resistor & Inductor (RL)	R-jpg R	R-jpg R	R-jpg L	R-jpg Cluster 1	R-jpg Cluster 1
	L-jpg L	L-jpg L	L-jpg R	L-jpg Cluster 1	L-jpg Cluster 2
Capacitor & Inductor (CL)	C.jpg C	C.jpg C	C.jpg C	C.jpg Cluster 1	C.jpg Cluster 1
	L-jpg L	L-jpg L	L-jpg R	L-jpg Cluster 2	L-jpg Cluster 1
Capacitor & Voltage Source (CV)	C.jpg C	C.jpg C	C.jpg C	C.jpg Cluster 1	C.jpg Cluster 1
	V-jpg V	V-jpg C	V-jpg C	V-jpg Cluster 1	V-jpg Cluster 1
Resistor, Inductor, Capacitor & Voltage Source (RLCV)	R-jpg R	R-jpg R	R-jpg R	R-jpg Cluster 1	R-jpg Cluster 1
	L-jpg L	L-jpg L	L-jpg R	L-jpg Cluster 2	L-jpg Cluster 1
	C.jpg C	C.jpg C	C.jpg C	C.jpg Cluster 3	C.jpg Cluster 2
	V-jpg V	V-jpg C	V-jpg C	V-jpg Cluster 3	V-jpg Cluster 2

Fuzzy C means, it is better to prefer K-means over Fuzzy C means. Thus we have the task of selecting the algorithm among the four machine learning algorithms. The

problem with unsupervised learning is, the output clusters are to be tagged manually for results which is not possible in our case. So we can stick to the three supervised learning algorithms, decision tree, SVM and K-NN. Decision tree has a major draw-back, it is not mandatory that a tree can be built for all classes based on their feature. So to reduce the risk of failure the algorithm can be skipped. SVM's as discussed has the limitation over the number of output categories. So for K components we wanted $K(K + 1)/2$ classifiers, which is very much time and cost consuming. Thus we will stick to the simplest and most efficient of the algorithms, K-NN. As K-NN is selected for component recognition, the same is applied for numbers, alphabets and symbol recognition which are comparatively simpler and has the similar process. The various algorithms were tested in about half-a-dozen of different circuits with various combinations of components and the accuracy of the outputs is represented as a graph below. The above results of various circuits with different components were compared, and their accuracies are as in Fig. 1.

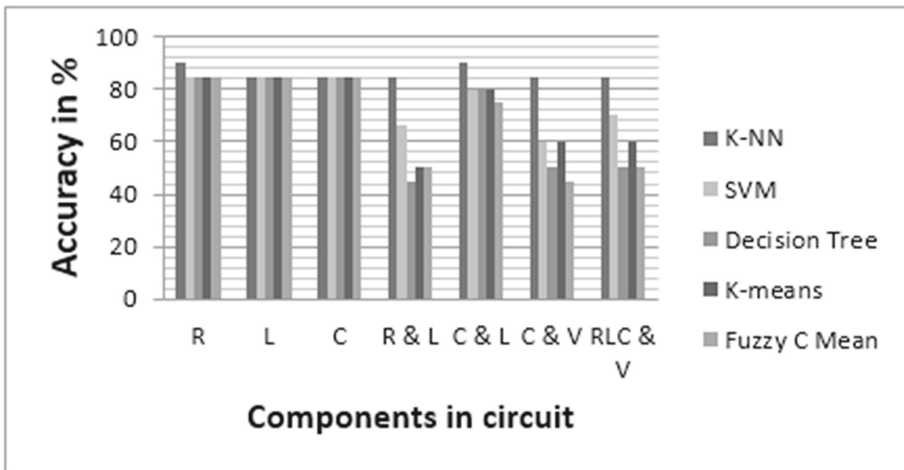


Fig. 1. Accuracy of machine learning algorithms for circuits with different components

8 Virtual Reality

The growth of virtual and augmented reality is really enormous and it is very appreciable as it is used many constructive applications [6–9]. Virtual Reality is a computer generated environment with realistic images and sound, where the user can interact virtually. The user’s presence is also simulated inside the environment for more realistic feel. VR also simulates all 360° of the environment in 3Dimensional space. For our project, a simple and very reliable hardware is selected. The LEAP motion sensor is used to simulate the virtual environment and interact with it. This sensor has the capability of capturing any hand gesture and simulate the same in the virtual environment to give the user a feel of interacting in the virtual space. The device has 2 cameras and 3 IR LEDs. They can track IR rays of wavelength 850 nm which is outside visible spectrum. The

sensors gets the IR based image from the LEDs, separated into two cameras. IR image only senses live objects (hands in here) and dual camera image give the depth information. Thus it simulates 3D gestures in the VR space.

Initially, we tried to generate a circuit by dragging and dropping the components from a tool kit into virtual space. This is achievable using the drag and drop gesture of the sensor. By default it is supported by the sensor. So, the virtual UI will have a range of supported components. The user will be prompted to drag each and every component and drop it into a small bucket at the bottom space. As the components are getting added, the circuit will be built in the background and finally it will be displayed. This works absolutely fine for simple circuits. But when parallel or complex circuits are considered, this approach fails. The problem is that a node will be connected to two or more components. This cannot be achieved with this UI. So a different approach is sought. The Leap sensor supports a different gesture called Pinch Draw. Here, the sensor detects a pinch gesture and captures all the motion of the pinch. Thus, the user can pinch and then draw the circuit. This need not mandatorily be continuous. It supports drawing in parts. For implementing this, we installed the Leap motion software with the help of the installer so that the leap motion controller is ready for usage. Subsequently a unity application is created to which three modules namely leap motion core asset module, Detection module and leap motion hand module are imported. With help of these modules a scene is created in which the circuits can be drawn using pinch gesture.

9 Proposed Solution

The architecture of the proposed solution is shared below in Fig. 2 and the same is discussed. The proposed solution is a mobile application with user friendly UI. The app has two phases, training phase and output generation phase. The initial setup of the app

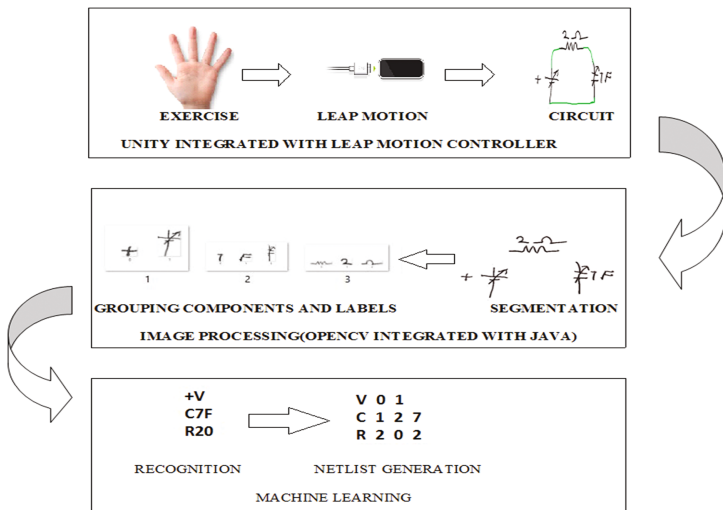


Fig. 2. Architecture diagram of the proposed solution

is the training phase. In the training phase, the app allows you to select specific components as shown in figure and then draw the same to add it to the training data. In the lateral phase, when a new circuit is given as input for NetList generation, the components are identified based on the samples drawn during the training phase. Thus, the training data is also populated by the user as shown in Fig. 3, increasing the efficiency of the final output to a great extent.

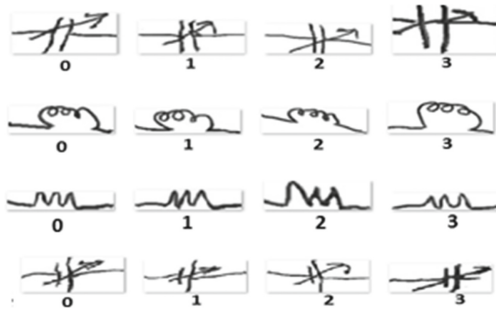


Fig. 3. Various components in training data

For generating a NetList, the user should select the NetList button on the first UI shown in Fig. 4. Immediately, the Virtual Reality mode gets triggered. Then the mobile is to be placed in the head mounted display (HMD) and connected to the external hardware that was discussed earlier. In the virtual reality environment, the UI has three buttons as shown in Fig. 5.

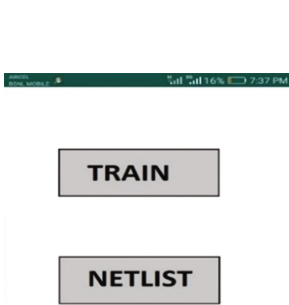


Fig. 4. Main menu UI



Fig. 5. Training phase UI

First the user should select the components button and draw all the components of the circuit along with the signs, values and units. Now the connecting wires are to be drawn. The user should click in the wire button and then draw the wire. This is done to change the color of the components and the connecting wires, to make it easy at the stage of component recognition which is discussed later. Finally, when the input circuit is over, the user should click on the save mode. Now, the app takes a screen shot of the circuit drawn and saves it in the phone. It is to be noted that all these buttons are in virtual

environment, and the clicks and drawing are also based on hand gestures. After saving the image, the application turns back to normal working, now the mobile can be disconnected from the sensor and removed from the HMD.

Then the new UI opens with a new set of buttons for next step. The steps include, component separation, component recognition, grouping and NetList generation as shown in Fig. 6. In component separation phase, each and every component is separately stored as an image. The system cannot separate the wires and components from the fully connected circuit. Whereas, the symbols and values can be separated as they are not connected with other components. A clustering algorithm runs recursively to cluster all the colored pixels connected together and surrounded by the white pixels. Once all the clusters are fixed, each cluster is treated as an individual component and saved as a separate image. Now the circuit alone is saved in a separate image. From this image the components are to be separated. First a raster scan algorithm is applied to the image, each image from the top left corner to bottom right corner is scanned and checked if it is colored. If the pixel is not black, then it is the connecting wire. All the colored pixels are removed, now the image has only the components separated from each other [5]. Then the same clustering algorithm can be applied for saving each component separately. Thus, component separation phase is complete. These are saved as contours, with their co-ordinates of the original image.

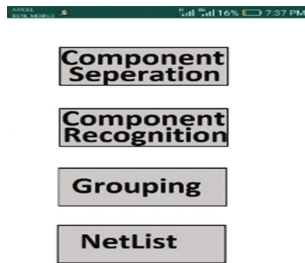


Fig. 6. NetList phase UI

In component recognition phase, each of the above separated components is applied to the machine learning algorithm discussed earlier. Thus each component can be identified. Then all the symbols and values are also identified. Once the components are identified, their file names are renamed correspondingly. Also, a count is maintained in the naming convention for multiple components of same type, like Capacitor1, Capacitor2, etc. to differentiate components. Then these components and their corresponding symbols are to be grouped. The grouping phase is done based on the co-ordinates from the original image. The components are treated as rectangles with co-ordinates available for all four corners. Grouping of the components is based on the distance between them. A component is grouped with the symbol and value with the minimum distance from the component. Once the grouping phase is over we will have a folder structure for each component. All these component folder structures are available inside a master folder. The final phase is the NetList generation. The algorithm traverses through each component and appends the name of the file into a text document. The name of the image file

is sufficient for the NetList as we follow a naming convention for differentiating each component separately. Followed by the component name is the starting and the ending node. The folder structure is organized in a way where, the components in the top comes first, middle comes next and the lower part follows. So, nodes can be numbered continuously like 0 1, 1 2, etc. till the end. The nodes will be followed by the value and symbol file names. Thus the Netlist is completed and saved in the specified location.

10 Experimental Results

On running the application, the circuit can be drawn with the help of leap motion controller and the corresponding netlist is generated. With a sample circuit given as input the following results were achieved. Figure 7 shows the drawing of the input I the virtual environment. Whereas Figs. 8, 9 and 10 shows the components drawn, completed circuit with wires and NetList generated correspondingly in order.

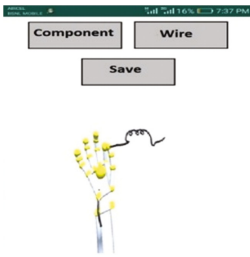


Fig. 7. Drawing in VR

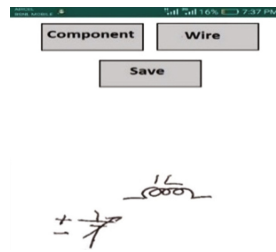


Fig. 8. Components drawn

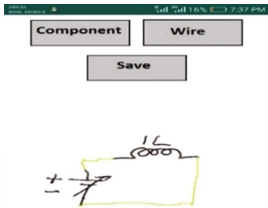


Fig. 9. Circuit completed



Fig. 10. NetList generated

11 Conclusion

Needless to say, there are some aspects for improvement and it can be incorporated in the future releases. The system can be made to suggest the user with previously drawn circuits, if the circuit being drawn is the same as one in the history. While this application is made for netlist generation, it would be extremely useful if VHDL or VERILOG code can be generated through this application. Usage of neural networks along with machine learning would be another aspect of this application. All these are at advanced levels

which would need more time and expertise for these incorporations, which will become possible in near future.

References

1. Bailey, D., Norman, A., Moretti, G.: *Electronic Schematic Recognition*. Massey University, Auckland (1997)
2. Feng, G.-H., Sun, Z., Christian, V.-G.: Hand-drawn electric circuit diagram understanding using 2D dynamic programming. In: *Proceedings of the 11th International Conference on Frontiers in Handwriting Recognition (ICFHR 2008)*, Concordia University, Montreal, Quebec, Canada, 13–15 August 2008
3. Palakodety, R., Shah, V.: *Hand-Drawn Circuit Recognition*. Massachusetts Institute of Technology, Cambridge (2005)
4. Sridar, S., Subramanian, K.: Circuit recognition using netlist. In: *2013 IEEE Second International Conference on Image Information Processing (ICIIP)*, IEEE (2013)
5. Vasudevan, S.K., Venkatachalam, K., Anandaram, S., Menon, A.J.: A novel method for circuit recognition through image processing techniques. *Asian J. Inf. Technol.* **15**, 1146–1150 (2016)
6. Sreedasyam, R., Rao, A., Sachidanandan, N., Sampath, N., Vasudevan, S.K.: Aarya—a kinesthetic companion for children with autism spectrum disorder. *J. Intell. Fuzzy Syst.* **32**(4), 1–6 (2017)
7. Vasudevan, S.K., Vivek, C., Srivathsan, S.: An intelligent boxing application through augmented reality for two users – human computer interaction attempt. *Indian J. Sci. Technol.* **8**(34), 1–7 (2015)
8. Sundaram, V.M., Vasudevan, S.K., Santhosh, C., Barath Kumar, R.G.K., Deepak Kumar, G.: An augmented reality application with leap and android. *Indian J. Sci. Technol.* **8**(7), 678–682 (2015)
9. Geethan, P., Jithin, P., Naveen, T., Padminy, K.V., Shruthi Krithika, J., Vasudevan, S.K.: Augmented reality X-ray vision with gesture interaction. *Indian J. Sci. Technol.* **8**(S7), 43–47 (2015)

FPGA-Based Heavy-Ion Beam Trajectory Estimation and Control for Superconducting RF Cavity Resonator Applications

B. Christopher¹, S. Kiruthika¹, S. Lakshmi¹, R. Mugunth Krishnan¹,
and G.A. Shanmugha Sundaram^{1,2,3(✉)}

¹ Department of Electronics and Communications Engineering,
Amrita School of Engineering, Amrita Vishwa Vidyapeetham,
Amrita University, Coimbatore, India
ga_ssundaram@cb.amrita.edu

² Amrita School of Engineering, Center for Computational Engineering
and Networking (CEN), Amrita Vishwa Vidyapeetham,
Amrita University, Coimbatore, India

³ SIERS Laboratory, ASE Coimbatore, Amrita University, Coimbatore, India

Abstract. The manipulation of heavy ion beams by applying fluctuating external electric and magnetic fields is discussed here. The critical beam parameters that are identified with beam trajectory are identified and their properties are discussed. Most of the beam measurements are based on the electromagnetic fields induced by the beam. The analog signals obtained from the sensors are amplified and shaped before they were converted into numerical values, which are then further treated in order to extract meaningful machine parameter measurements. The primary goal is to calculate the ion beam position and degree of spatial coherence. Beam position monitors (BPM) are used for this purpose. The pickup (PU) points were defined to collect the horizontal and vertical difference signals, that in turn were used to calculate the position and trajectory. A system of quadrupoles, modified quadrupoles and octupoles was replacably used for confining and focusing the beam. The BPM readings were modeled based on the readings from simulated ion motion in MATLAB. These results were further processed with noise and digitized and fed into an FPGA implementation of the Kalman filter for estimation of the ion trajectory.

Keywords: FPGA · Ion beam diagnostics · Heavy ion trajectory · Kalman filter · Quadrupole · Octupole

1 Introduction

High currents of heavy ions beams, such as that of Niobium (Nb), have large values of potential due to the large positive ion charge they carry. A strong, externally applied electric and magnetic field is required to control the beam, in terms of its confinement and forward propulsion. The beam cross section tends to grow as a result of the radial motions of ions arising from repulsions due to space charge effects [1] between them. This in turn causes severe loss to the critical quality parameters of the beamed ions [13].

To ensure proper focusing of the beam, suitable control system needs to be employed. Beam diagnostics is one of the major stages involved in the monitoring, control and collimation of the ion beam [5, 6, 11, 14]. The data obtained from the diagnostics unit can be processed and accordingly the electro-magnetic fields can be altered, to bring about the desired ion trajectory and density.

Towards achieving better quality metrics for the Nb-ion beam, the best combination of electric and magnetic fields that are required for its propagation have to be determined. Also an estimation of the trajectory of the ion beam was done using a combination of simulations that culminates in its FPGA implementation.

An electric quadrupole was used to generate the electric/magnetic field for confining and axially propagation of the high energy heavy ion beam over a certain distance with minimum deviation, without any significant loss to its energy. Octupoles and modified quadrupoles perform about the same function of quadrupole, in terms of their ability to confine the beam and ensure a forward propagation, but with better results. A velocity distribution that follows a Gaussian, normal pattern has been assumed for the Nb ions, as they issue out of the thermionic source located at one end of the cylindrical column that is evacuated to UHV levels. Singly charged Nb ion is chosen as the heavy ion for generating the beam since it finds extensive use in specific liquid-metal based nuclear applications and in the fabrication of various rocket components [15]. It also has excellent superconductive properties, finding use as thin layers in radio frequency resonators. Nb offers good resistance to corrosion, and also has a larger atomic mass.

Particle tracking algorithms such as the Kalman filter [3] are used for the best estimation of the parallel trajectory of ion beam. The Kalman filter technique was developed to determine the trajectory of the state vector of a dynamical system from a set of measurements taken at different times [5, 7]. The Kalman filter proceeds progressively from one measurement to the next iteratively, improving the knowledge about the trajectory with each new measurement, and is hence has been chosen so as to obtain a progressively refined estimate for the beam trajectories under faster convergence [6].

2 Methodology

2.1 Description

For the purpose of beam control, a replaceable system of electrical quadrupoles, modified electrical quadrupoles, and electrical octupoles was employed. An ideal quadrupole is a system of four conducting rods which are orthogonal in their orientation and extending to infinity. The length of each of the rods is 0.4 m. The modified quadrupole is a system in which two quadrupoles were aligned one above the other but with a shift of 45° in orientation. The length of each of the rods in this case is 0.175 m and the spacing between two quadrupoles is 0.05 m. The voltage applied to the rods is 1 V each.

2.2 Algorithm for Ion Motion Simulation

Step 1: Solving the Laplace equation to obtain a grid of voltage

Step 2: Calculate the electric field in x, y and z axes from the potential defined to the rods

- Step 3: Axial and radial magnetic values calculated from existing data
- Step 4: Interpolate the magnetic values to the required region
- Step 5: Load the electric and magnetic field values.
- Step 6: Define the parameters for the particle [mass, charge, velocity, frequency of AC] and convert to S.I. units.
- Step 7: Set the initial conditions such as position, angle, velocity, start and end time.
- Step 8: Solve the Lorentz equation
- Step 9: Plot the poles and ion trajectory for single particle.

Mathieu’s equations are used to define the motion control of Nb ions in the quadrupolar environment. It is a second order linear differential equation which serves a great deal in generating the stability curves, the functioning of the quadrupole and degrees of parameter dependencies [2].

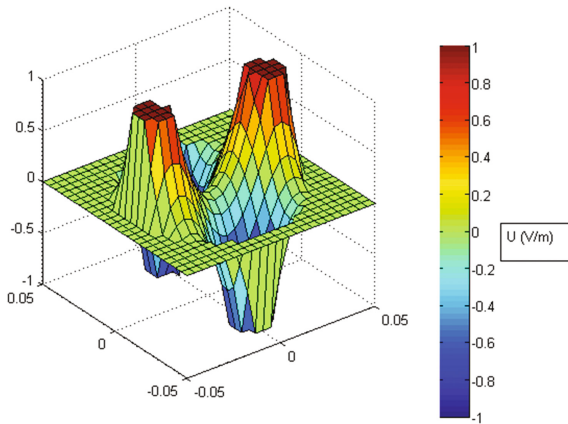


Fig. 1. Distribution of electric fields due to a quadrupole (in units of V/m)

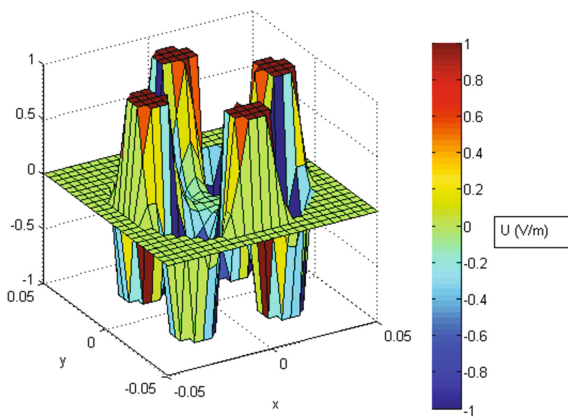


Fig. 2. Distribution of electric fields due to an octupole (in units of V/m).

Potentials are applied suitably on the conducting rods in a specific manner to achieve the desired results. A combination of periodically fluctuating and bias voltages are applied to the rods such that the adjacent rods have opposite polarities. This is done in order to provide a strong electric field which helps in the propulsion of the ion beam with minimum deviations. The same procedure was carried out for an octupole (Fig. 1) and a modified quadrupole as well. A comparative study between quadrupoles and octupoles (Fig. 2) has been performed from such efforts.

A modified quadrupole and a single set of quadrupoles was designed in computer models and their behavior was studied in a simulation exercise. A comparative study between a modified quadrupole and a single set of quadrupoles has been done. Generally quadrupoles with hyperbolic surfaces are used in mass filters because they provide greatest mass resolution for minimum power [2]. Since this project does not deal with mass filtering, mass resolution parameter can be neglected. Thus cylindrical rods are used for modeling octupoles because they are the best suited for RF cavity resonator applications.

Ion motion simulation was done inside the structure of the quadrupole to guarantee that the modified quadrupole works appropriately. It was modeled considering the electric field created by the rods of the quadrupole and their induced magnetic field. The differential equations for the ideal quadrupole case and for the ion motion in general, from fields generated by potential grids, were solved in each direction using a MATLAB solver called Ode45 [9]. The quadrupole is able to affect the ion motion because of the electric and magnetic fields produced by the rods. Hence the electric fields are now modeled from the generated potential grid values. Electric fields are directional and hence the field values are calculated separately in each direction.

The magnetic fields produced by the electrodes are calculated from data obtained by experiments [2]. The effect of the magnetic field on the ions can be modeled using the well-known Lorentz equation [10]. The initial values of the ions are assigned properly for the efficient functioning of the differential equations involved in tracing the trajectory of the ion [12]. A separate function was created where the electric and magnetic field values are interpolated and used in the equations of ion motion. The results of this function are in turn applied to the differential solver Ode45 to obtain the actual position of the ion in motion. The path of the ion was plotted along with the modified quadrupole and a single quadrupole environment where the particle is propagating. Similarly the design of ideal quadrupole was extended to octupole for the same length. The electric field produced by the octupole has more circular profile in cross-section, and a better degree of symmetry than that of the field generated by the quadrupole. Evidently, an octupole provides more optimal results than a quadrupole.

2.3 Kalman Filter Implementation

A field programmable gate array (FPGA) module was incorporated in the beam diagnostics exercise in order to acquire data from the beam position monitors (BPM) and processes it to remove typical noise signatures present in the signals received. The BPM data are modeled based on the readings from simulated ion motion in MATLAB. This simulated data is further treated with noise and digitized and then fed to Kalman filter block for the estimation of trajectory. The Kalman filter has a wide

variety of applications in the estimation and noise filtering domain. The lower power utilization of the Kalman filter FPGA implementation makes it an ideal choice for the noise filtration of the trajectory data. The essential parameters involved in the Kalman filter design are governed by the following set of equations [8]:

For the computation of Kalman Gain:

$$K_k = P_k^- H^T (H P_k^- H^T + R)^{-1} \quad (1)$$

For obtaining the update estimate with measurement Z_k :

$$x_k = x_k^- + K_k (z_k - H x_k^-) \quad (2)$$

In most of the particle beam line facilities FPGAs are used for acquisition and estimation of real-time beam parameters [14, 16]. Therefore the above mentioned Kalman filter was programmed in the VHDL language. A set of constraints were defined. With the help of these definitions and the VHDL code, the design was synthesized. An FPGA board of Kintex-7 family [17] was employed for implementing the Kalman filter.

3 Results and Discussion

3.1 Analytical Simulation for Beam Diagnostics

The motion of the Nb ion has been simulated in the MATLAB environment. The path of ion inside an ideal quadrupole, a modified quadrupole and an octopole setup was also simulated in the MATLAB and analyzed. Since the heavy ion has greater inertia, the trajectory of the ion tends to follow a rectilinear path despite the effect of ponderomotive forces on the ion, as shown in Figs. 3(a) and (b).

The electromagnetic fields help in confining and propelling the ions in the system. The trajectories of Nb ion were studied under different focusing setups.

Comparatively the octupole environment provided better confinement results for the beam traversal (Fig. 3c). Evidently, with the addition of more pairs of electrical poles, the ion beams exhibit a trajectory that is progressively more symmetric and collimated.

The fields of the poles tend to have a more conspicuous effect in terms of confinement on the ion in the case of octopoles than in quadrupole. While the modified quadrupole gives a better confinement than a single quadrupole, and this is superseded by that with an octupole. This influence of fields can be observed from the Table 1, which gives an overview of the average deviation of the ions from the center of the beam in the simulated environments.

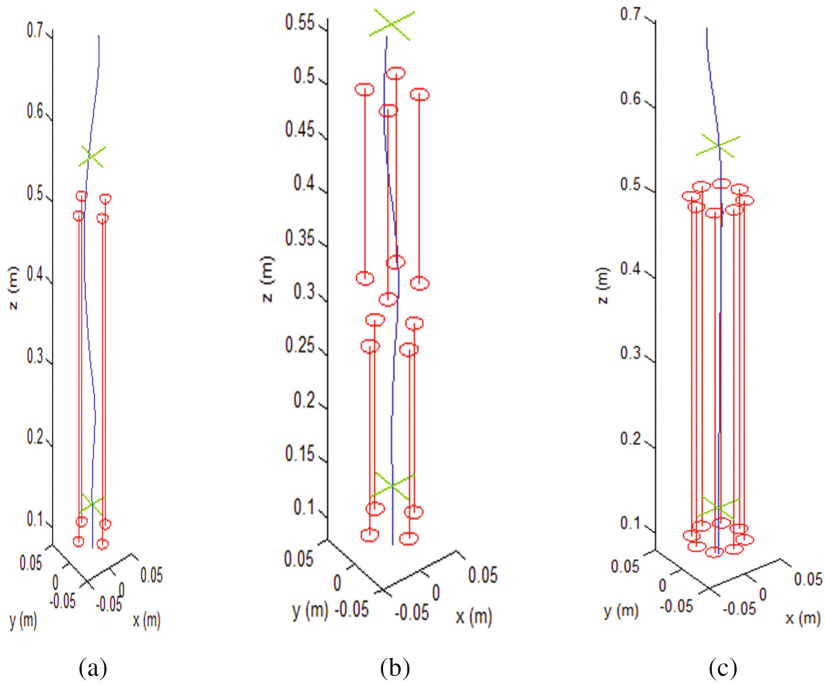


Fig. 3. (a) Trajectory of Nb ion in a quadrupole. (b) Trajectory of Nb ion in a modified quadrupole. (c) Trajectory of Nb ion in an octupole

Table 1. Average deviation of Nb ions from the axis of the evacuated cylindrical column

Poles	Average Deviation of Nb (10^{-2} m)
Quadrupole	0.0121
Modified Quadrupole	0.0116
Octupole	0.0091

The results observed show that the modified quadrupole and the octupole were 4.13% and 24.79% better in confining the Nb ions than an ideal quadrupole, respectively.

In Figs. 4, 5 and 6, with the x and y orientations as the direction of electric field and magnetic field correspondingly, the ion propagation is in along the z direction (as seen head on). The Kalman filter gave the best estimates of the position of the ion which is the crucial stage in beam diagnostics. Process noise variance and the measurement noise variance were defined to be 0.63681 and 0.000649, respectively, based on white Gaussian noise profile in the Kalman filter simulation. The Kalman filter closely followed the original trajectory and convincingly estimated the path of the ion beam, as can be inferred from Fig. 7.

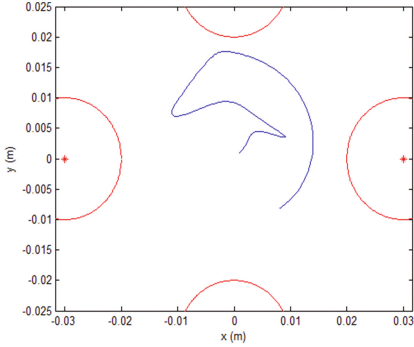


Fig. 4. Ion trajectory of Nb in a quadrupole (head on view).

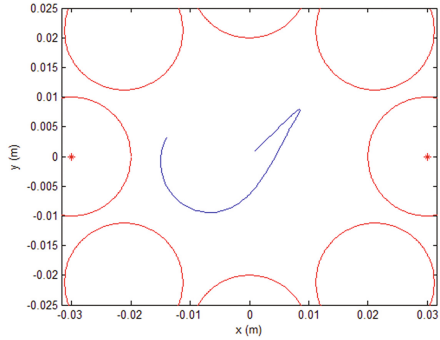


Fig. 5. Ion trajectory of Nb in an octupole (head on view).

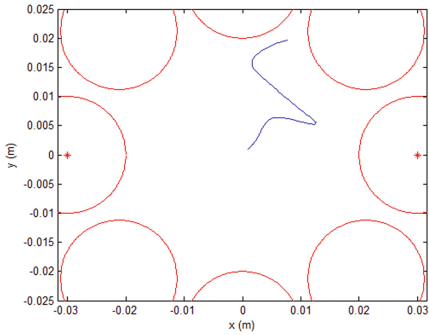


Fig. 6. Ion trajectory of Nb (top view)

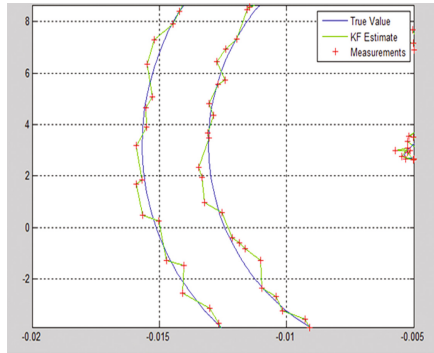


Fig. 7. A 2D Tracking of Nb ion using the Kalman filter simulation block

3.2 FPGA Simulations

The Kalman filter was then programmed in the VHDL language as part of the FPGA implementation. Xilinx ISE Suite was used for the design synthesis. The Kalman filter involved some division operations which needed higher degree of precision and coding, and such operations in VHDL was challenging. Hence a core generator module available in the ISE Suite was used to generate an optimal division code which was then incorporated into the original Kalman code. The VHDL code was then taken to the design phase where the timing and routing constraints were specified.

The top level schematic of Kalman filter in FPGA is depicted in Fig. 8. Kalman filter designed uses 16-bit values. The analog to digital converter (ADC) outputs from the BPM, which are modelled in MATLAB, was fed to the z_in port. The P_in ports receive the values of covariance matrix which are constantly updated through the feedback provided from the P_out ports. The estimates from the filter are drawn from

x_est ports. The x_in ports are used for feeding the delayed version of the previous estimates which are needed for efficient Kalman filter operation. The initial values of the feedback inputs, *i.e.* x_in , and the P_in are pre-defined within the code.

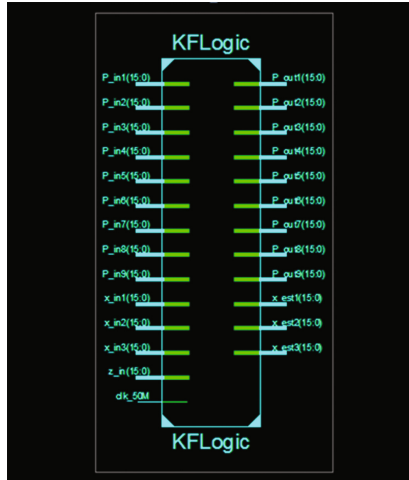


Fig. 8. Top level FPGA design implementation

Synthesis and performance evaluation was subsequently performed, and the corresponding metrics were exported. The outputs from the MATLAB program were used as the test inputs to check the working of the FPGA. Results from MATLAB and the FPGA were found to be comparable. The post-synthesis reports on device utilization and thermal summary were then generated (Figs. 9 and 10) in order to demonstrate the optimal design that uses various gates, latches and interconnects. A total of 2318 slice registers had been used out of 407,600 that amounts to upto 1% of the available registers in Kintex-7.

Thermal Summary	
Effective TJA (C/W)	1.8
Max Ambient (C)	84.7
Junction Temp (C)	25.3

Fig. 9. Thermal summary of the FPGA implementation of the Kalman filter

Device Utilization Summary:

Slice Logic Utilization:

Number of Slice Registers:	2,318 out of 407,600	1%
Number used as Flip Flops:	2,253	
Number used as Latches:	0	
Number used as Latch-thrus:	0	
Number used as AND/OR logics:	65	
Number of Slice LUTs:	2,452 out of 203,800	1%
Number used as logic:	2,252 out of 203,800	1%
Number using 06 output only:	1,148	
Number using 05 output only:	18	
Number using 05 and 06:	1,086	
Number used as ROM:	0	

Fig. 10. Device utilization summary for the FPGA implementation of the Kalman filter

4 Conclusions

A suitable combination of electric and magnetic fields was designed for the proper confinement of the ion beam and its forward propagation. The simulation of the Nb ion beam requires to take into account the electric and magnetic fields induced by the high current beam profile and its interaction with the EM field of the quadrupole, modified quadrupole and octupole. Taking into considerations these interactions and the large number of ions within the beam, the programming complexity had correspondingly increases. The factors affecting the motion of a heavy ion carrying high current were understood and included in the simulations. Usage of octopoles provides better collimation than the traditional approaches. The improved results could find extensive utilization in the application of ion beams, towards generating focused high currents, suitable for obtaining uniform thin film coatings in superconducting RF resonator cavities, compared to traditional approached such as CVD. Others parameters like the orientation, and RF voltages can be tweaked to obtain varied result. The deviations of the particle from its initial trajectory were found to be within tolerable limits. Results from the FPGA implementation of the Kalman filter showed its remarkable ability to closely track the original propagation path and hence estimates the trajectory of ions.

References

1. Goncharov, A., Protsenko, I., Yushkov, G., Brown, I.: Manipulating large-area, heavy metal ion beams with a high-current electrostatic plasma lens. *IEEE Trans. Plasma Sci.* **28**(6), 2238–2246 (2000)
2. Giraud, K.M.: Simulation and manufacture of a quadrupole mass filter for a Be7 ion plasma. Bachelor of Science. Brigham Young University (2008)
3. Kalman, R.E.: A new approach to linear filtering and prediction problems. *J. Basic Eng.* **82**, 35 (1960)
4. Welch, G., Bishop, G.: An Introduction to the Kalman Filter, vol. 1st ed, pp. 21–24. ACM SIGGRAPH, Los Angeles, CA (2001)

5. Manoj, S., Shanmugha Sundaram, G.A.: Dynamic spectrum of long wave terrestrial radio signals during episodes of ionospheric disturbances caused by solar activity. *J. Chem. Pharm. Sci.* **9**(1), 548–553 (2016)
6. Kasproicz, G.: Determination of beam intensity and position in a particle accelerator. Ph. D. Thesis, Warsaw University Of Technology (2011)
7. Fruhwirth, R.: Application of Kalman filtering to track and vertex fitting. *Nucl. Instrum. Methods* **262**(2–3), 444–450 (1987)
8. Baliyan, J., Aggarwal, A., Kumar, A.: Implementation of Kalman Filter using VHDL. *Int. J. Sci. Eng. Technol. Res.* **03**(08), 1569–1575 (2014)
9. MathWorks: MATLAB. www.mathworks.com
10. Cerati, G., Elmer, P., Lantz, S., McDermott, K., Riley, D., Tadel, M., Wittich, P., Würthwein, F., Yagil, A.: Kalman filter tracking on parallel architectures. *J. Phys.* **664**(7), 072008 (2015)
11. Rashmi, G., Shanmuga Sundaram, G.A.: Study of lower D-region of ionosphere from VLF signal perturbations. *J. Chem. Pharm. Sci.* **9**(1), 591–593 (2016)
12. Bai, G.: Modeling and experiments on injection in to University Of Maryland Electron Ring. Master of Science Thesis, Graduate School of the University of Maryland, College Park (2005)
13. Goncharov, A.: The electrostatic plasma lens. *Rev. Sci. Instrum.* **84**(2), 021101 (2013)
14. Cocq, D., Jensen, L., Jones, R., Savioz, J.: First beam tests for the prototype LHC orbit and trajectory system in the CERN-SPS. In: DIPAC, pp. 207–209 (2001)
15. Heisterkamp, F., Carneiro, T.: Niobium: Future possibilities – technology and the market place. In: *Niobium Science and Technology, Minerals, Metals and Materials Society, Metals and Materials Society Minerals* (eds.): Proceedings of the International Symposium Niobium 2001, Niobium 2001 Ltd, Orlando, Florida, USA, ISBN 978-0-9712068-0-9 (2001)
16. Kline, D.M., Ross, S. K.: Employing RTEMS and FPGAs for beamline applications at the APS. In: Proceedings of the PCaPAC, pp. 27–29, Saskatoon (2010)
17. Kintex-7 FPGA Family – Xilinx. www.xilinx.com/products/silicon-devices/fpga/kintex-7.html

PAM4-Based RADAR Counter-Measures in Hostile Environments

S. Srivatsa¹ and G.A. Shanmugha Sundaram^{1,2(✉)}

¹ Center for Computational Engineering and Networking (CEN),
Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham,
Amrita University, Coimbatore, India
ga_ssundaram@cb.amrita.edu

² Department of Electronics and Communications Engineering,
Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham,
Amrita University, Coimbatore, India

Abstract. Signal jamming and counter-jamming have become areas of intensive research in the field of radar systems engineering. Either of signaling systems and jamming techniques have had concomitant development. While non-linear frequency modulation (NLFM) signals and polyphase codes continue to be the most preferred radar signals, they still are very much susceptible to jamming, and radars could be easily deceived, given that these signals are predictable in nature. In the work reported here, radar signaling based on 4-level pulse amplitude modulation (PAM4) is considered as an effective modulation scheme to evade jamming, since inclusion of PAM4 in typical pulsed radar signals shows a marked improvement as a function of target detectability. In recent years, PAM4 has gained prominence having had a demonstrable 100+ Gbps of data transmission rate over optical fibers. PAM4 is evaluated against conventional modulation schemes such as NLFM and QAM techniques in terms of burn-through range, cross-over range, and bit error rate. Simulation is done with a multiple-input multiple-output (MIMO) transmit-receive system that are used to replicate a bi-static radar configuration. Various criteria that are typical of a hostile signal environment such as multipath propagation and fading, atmospheric effects, slow fading due to target motion, and also a self-screening jammer are considered for evaluation of performance of PAM4. It is observed that PAM4 performs much better than conventional radar signals in the presence of a jammer signal, thereby indicating its potential as an effective counter-measure in hostile signal environments.

Keywords: PAM4 · MIMO tracking radar · BER · Jamming · Burn-through range · Cross-over range

G.A. Shanmugha Sundaram—SIERS Laboratory, ASE Coimbatore, Amrita University, India.

1 Introduction

Radar has been the backbone in surveillance systems from many decades, especially in the field of air traffic management and oceanic transportation. The science behind radar is all about the total and relative power reflected off a target, which is a function of the size and shape of the target object, the range, surface material of the target, and the channel through which the wave propagates. Recent developments consider the usage of multiple-input-multiple-output (MIMO) antennas to be a vital technique in futuristic radar technology [1]. MIMO is a highly successful technology in wireless communication engineering and has intensively attracted research in the field of radar [2].

Two important configurations in MIMO radars may be observed in the literature. One considers multiple antennas that are spatially separated to exploit the spatial diversity in the observation of a target. This configuration has proven to be more efficient in parametric estimation of a target. Second type is a antenna configuration where multiple transmitters are co-located to make use of the waveform diversity, which is possible with antenna arrays. This configuration improves the angular resolution, thus providing more detail on the target surface and also allows distinguishing multiple targets that are closely spaced [3].

Surveillance in airborne and marine environments are mainly affected by undesired reflections off the land and sea surfaces [4]. These high intensity reflections mask the actual target from being detected, and are collectively called as clutter. Space-time adaptive processing (STAP) plays an important role in improving performance of moving target detection radars [5,6].

Radar technology is not fool-proof. Just after the development of radar, researches started coming up with ideas to make objects invisible to a radar. Some of the earliest ideas were to design the surface of the fighter jets such that the total radar cross-section (RCS) of the target reduces. Further developments included usage of surface materials that absorb the signals radiated by the radar transmitter. But, all these ideas were effective on particular wavelengths of electromagnetic spectrum. Nowadays, more sophisticated techniques like signal jamming and deception are being used. Development of Digital Radio Frequency Memory (DRFM) has made aircrafts very smart in self-screening themselves from almost any type of radar. This has started a new age of electronic warfare, where intelligent electronic systems are used to exploit the hostile use of the electromagnetic spectrum [7].

The signaling scheme used in radar thus plays an important role in the quality of detection of a target. Various signals commonly used in radar systems are linear frequency modulation (LFM), non-linear frequency modulation (NLFM), Barker codes, polyphase codes, etc. [8]. These signaling systems are based on a technique called pulse compression, that aims at improving the range resolution without compromising the energy transmitted per pulse. In this work, pulse amplitude modulation with 4 amplitude levels (PAM4 or 4-ary PAM) is considered as a signaling scheme for radar systems. The quality of this modulation system is compared against other pulse compression techniques and quadrature amplitude modulation - 16 levels (QAM16). Also, the effectiveness of PAM4 in

the presence of jammer is analyzed with the help of achievable burn-through range and cross-over range for a given signal-to-interference ratio (SIR).

The paper has been organized as follows: Sect. 2 describes theoretical background with their mathematical description for PAM4, MIMO radar and jamming; Sect. 3 gives a detailed description of the methodology of the work incorporated; Sect. 4 provides the results obtained. Also, the inferences made out of the results are showcased. Finally, conclusions and possible enhancements to the work are to be found in Sect. 5.

2 Theoretical Background

2.1 4-Level Pulse Amplitude Modulation

With an exponential increase in the amount of data generated and processed everyday, there is a requirement for similar improvement in the data transmission. Data rate achievable is directly dependent on the amount of bandwidth being used. But, bandwidth cannot be increased indefinitely due to many implications like frequency sharing among various services, cost, and various environmental factors. Also, lower bandwidth ensures lesser losses for a given center frequency and reduces the PCB circuit traces required [9].

Pulse amplitude modulation (PAM) is a signal modulation technique, where the data is contained in the amplitude information of the carrier signal. The generic form of PAM in digital modulation system is known by the name M-ary PAM, where M stands for the number of amplitude levels, which is a function of 2^N , $N = 1, 2, 3, \dots$. Increase in the value of M improves the bandwidth utilization but at the expense of signal-to-noise ratio (SNR).

For an M-ary PAM system, symbol error probability $P\{E_s\}$ is formulated as [10]

$$P\{E_s\} = \frac{M-1}{M} \cdot \text{erfc} \left[\sqrt{\frac{3}{M^2-1} \times \frac{E_s}{N_0}} \right] \quad (1)$$

and the bit error probability $P\{E_b\}$ is $P\{E_b\} = (1/k)P\{E_s\}$, where, k is the number of bits transmitted per symbol, i.e. $k = \log_2 M$. From the equations of $P\{E_s\}$ and $P\{E_b\}$, the required signal-energy-per-symbol to noise-power-spectral-density ratio (E_s/N_0) can be calculated. In the case of pulsed radar system the value of E_s/N_0 may also be represented by [10]

$$\frac{E_s}{N_0} = \frac{P_{avg} \cdot T_p}{N_0} = \frac{P_{sd}}{N_0} \times \frac{B_s}{B_p} \quad (2)$$

where, P_{avg} is average received power, T_p is the pulse repetition period, P_{sd} is the average power spectral density, B_s is the equivalent occupied bandwidth and $B_p = 1/T_p$ is the pulse repetition frequency.

PAM4 is a compromise between bandwidth requirement and SNR. PAM4 reduces required bandwidth to half for the same data rate, when compared to binary modulation techniques. But obtainable SNR is reduced by thrice. One important issue with PAM4 is that they suffer from timing skew. This effect may

be cleared observed on an eye diagram, which makes the eyes more narrower, further degrading the possibility of detection of transmitted data.

But, techniques like FFE (Feed Forward Equalization) in transmitter, DFE (Decision Feedback Equalization) and CTLE (Continuous Time Linear Equalization) in receivers have been successfully used to improve eye openings. Data rates of over 1 Gbps in wireless visible light communication (VLC) and over 100 Gbps in optic fiber cables (OFCs) have been demonstrated using PAM4 [11]. PAM4 is also more ISI resilient for a given data rate on lossy electrical channels [12].

2.2 MIMO Radar

Consider a radar system with MIMO antennas having M elements in transmitter and N elements in receiver [3]. Let the transmitted waveforms out of the M transmitting array be represented by a $M \times 1$ vector-

$$\phi(t) = [\phi_1(t), \phi_2(t), \dots, \phi_M(t)]^T \quad (3)$$

Each of the element of this vector, ϕ_i are orthogonal to each other. Let us consider that P sources of reflections are observed, that includes both desired target reflections via multipath propagation and unwanted background clutter. Then, the signals received at the receiver array may be expressed as $N \times 1$ vector-

$$y(t, \tau) = \sum_{l=1}^L \alpha_l(\tau) [a^\tau(\theta_l) \phi(t)] b(\theta_l) + z(t, \tau) \quad (4)$$

where, τ represents slow-time, $\alpha_l(\tau)$ and θ_l are reflection co-efficient and spatial angle of the l^{th} source. The parameters $a(\theta)$ and $b(\theta)$ are the steering vectors of transmitting and receiving antenna arrays. z is zero-mean AWGN term. The received signals are passed through a matched filter that leads to-

$$\tilde{y}(\tau) = \text{vec} \left[\int_{T_p} x(t, \tau) \phi^H(t) dt \right] \tilde{y}(\tau) = \sum_{l=1}^L \alpha_l(\tau) [a(\theta_l) \otimes b(\theta_l)] + \tilde{z}(\tau) \quad (5)$$

where, \otimes denotes the Kronecker product. Thus, the signal transmission with statistical MIMO can be designed to obtain desirable spatial properties for a radar [13]. They also have improved stability in target detection [1].

In a MIMO radar, clutter is more severe due to the addition of fast fading or the multipath propagation induced by multiple antenna elements. Sophisticated techniques for the reduction of clutter have been developed. STAP (space-time adaptive processing) is one such technique and may be used in MIMO radars [18].

2.3 Radar Jamming

Jammers are characterized as barrage jammers and deceptive jammers. Barrage jamming is the transmission of a high energy signal over a wide frequency range so that target reflection is completely masked [7]. Deceptive jamming is

an intelligent technique where the jammer misguides the radar receiver by various electronic and computational techniques. Advantage of barrage jammer is that it's easier to build and cover a wide portion of EM spectrum. But, deceptive jammers need less transmission power and are smart enough to mislead the radar.

Let us assume that there are J jammers and transmit their signal towards the radar, then the signal in Eq. 5 becomes [3]-

$$\tilde{y}(\tau) = \sum_{l=1}^L \alpha_l(\tau)[a(\theta_l) \otimes b(\theta_l)] + \sum_{j=1}^J \beta_j(\tau)[1_M \otimes b(\theta_j)] + \tilde{z}(\tau) \quad (6)$$

where, $\beta_j(\tau)$ and θ_j are the transmitted signal and spatial angle of j^{th} jammer. 1_M is a vector of 1's with length M . Considering $v(\theta_l) \doteq a(\theta_l) \otimes b(\theta_l)$ and $\tilde{v}(\theta_j) \doteq 1_M \otimes b(\theta_j)$ as the steering vectors of the (l^{th}) target and j^{th} jammer. Then, the Eq. 6 may be written as-

$$\tilde{y}(\tau) = \sum_{l=1}^L \alpha_l(\tau)v(\theta_l) \sum_{j=1}^J \beta_j(\tau)\tilde{v}(\theta_j) + \tilde{z}(\tau) \quad (7)$$

Effectiveness of jamming against a radar system depends upon various factors, like radar transmitter and jammer power, range, operating frequency and bandwidth, propagation and system losses, antenna pattern, physical properties of signal, signal coding and encryption, etc. Performance of radar signaling against jamming is a critical criteria to be considered while designing a radar system.

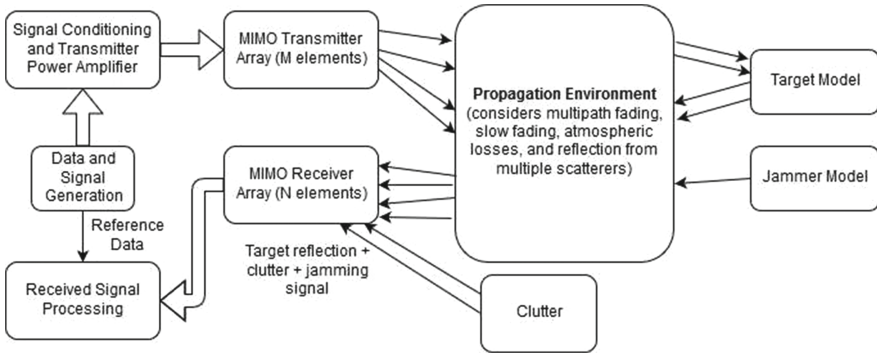


Fig. 1. Block diagram of the simulation of a radar with a jammer

In the case of MIMO radars, the sources of interference can appear in various forms and can have widely varying implications. Terrain-scattered jamming is a major impairment that occurs when jamming signals are reflected by the ground in a dispersive manner. So, a jammer's signal appear at the receiver array as a distributed source. The severity increases when signal direction from target reflection and jammer are the same, as in the case of self-screening jammers.

3 Methodology and Simulation

The block diagram in Fig. 1 shows an typical radar model used in the simulations discussed here. The transmitted pulses have been modulated with a PAM4 signal, instead of the bare high energy pulse. The modulated signal is then power amplified to boost the signal energy. A 4-array MIMO transmitter and a 4-array MIMO receiver are used as the antenna system [16]. The MIMO channel can be modeled as discussed in Sect. 2.2. MIMO transmitter transmits the same data with multiple antennas. This increases the net transmitted power and also enables diversity in the signals transmitted. The MIMO receiver acquires reflections from the target as well as from the background. The background reflections are termed as clutter. In the presence of an active jammer, the receiver detects the jamming signal too. The function of the signal processor following the receiver is to distinguish the target echoes from the undesirable clutter and jamming signals.

The propagation path is modeled as a multipath element. Multiple paths in the propagation of the signal is caused by the multiple scatterers present in the environment, which reflect the signals in different directions. A moving target also introduces slow fading to the transmitted radar signal. Atmospheric effects such as line-of-sight path loss, rain, fog, humidity, temperature and ionospheric effects are also considered in the propagation channel.

Three important models need to be addressed before simulating a radar system. They are the target model, clutter model and jammer model. Target model consists of properties like target RCS (radar cross section), range (distance between radar and target), velocity of target, direction of motion of target and the incident angle as viewed from radar. RCS depends mainly on the size, shape and surface material of the target. Some of the typical RCS values for various targets are listed in [20].

To model a jammer, the jammer's peak transmitter power, range of jammer from radar, its position, and velocity need to be considered. A jammer may be modeled as a self-screening jammer or as a stand-off jammer. The basic model of a jammer is described in Sect. 2.3.

A suitable clutter model must also take into account the type of radar deployed. Radars can be terrestrial or airborne or even sea-based. Terrestrial radars must consider reflections from ground, trees, buildings and any other terrestrial obstructions.

Some important design specifications like operating range, operating frequency, target RCS, maximum transmit power, pulse width, detection probability, etc. were obtained from the *Texas Instruments Designs* document [17] and from [18]. Some of the important design considerations for radar systems are pulse width, detection threshold, transmit power, gain of the antenna, targeted maximum radar range and jammer power. Pulse width affects many factors like maximum pulse energy, pulse repetition frequency, range of radar and the maximum velocity of the target that can be detected. It defines the transmit and dwell times [17]. Greater the pulse width, lesser the range observable by the radar and lesser would be the target velocity detectable. But, lower pulse width reduces

maximum pulse energy, thereby decreasing the SNR (signal-to-noise ratio) of received echoes. Again, lesser SNR reduces maximum range of radar.

Detection threshold is another such design parameter. It defines the minimum power of received echo to identify that the echo was from a target. Any power level below this threshold will be considered as clutter. These constraints must be addressed with proper choice of pulse width and also the appropriate modulation technique. Pulse compression is a good choice for overcoming these issues. A reference model, like the one presented in [17] would help design a simulated radar system.

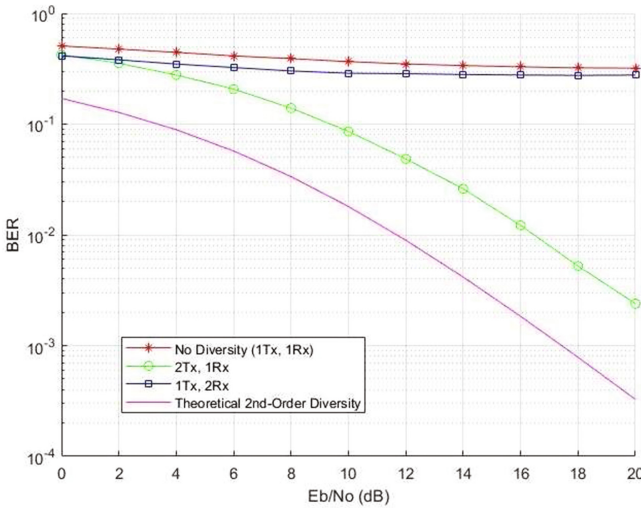


Fig. 2. BER performance of PAM4 for different MIMO transmit-receive antenna combinations

4 Results and Discussion

The performance of PAM4 is evaluated with some important criteria like BER, burn-through and cross-over ranges. BER is calculated with the help of Eq. 1. The simulation is done with various transmitter-receiver array combinations like, 1×1 , 1×2 , 2×1 , 4×1 , 1×4 and 4×4 arrays. Obtained BER performances are compared with QPSK (quadrature phase shift keying) and BPSK (binary phase shift keying) modulation techniques. A comparison of BER as a function of symbol-energy per noise-spectral density ratio for various MIMO transmit-receive antenna combinations is shown in Fig. 2. It is observed from the plots that a 4×4 antenna performs much better than other antenna configurations. In addition, a 4×4 antenna has been found as the most optimal combination in a MIMO system, and is also the choice in wireless communications systems such as the WLAN [16, 19].

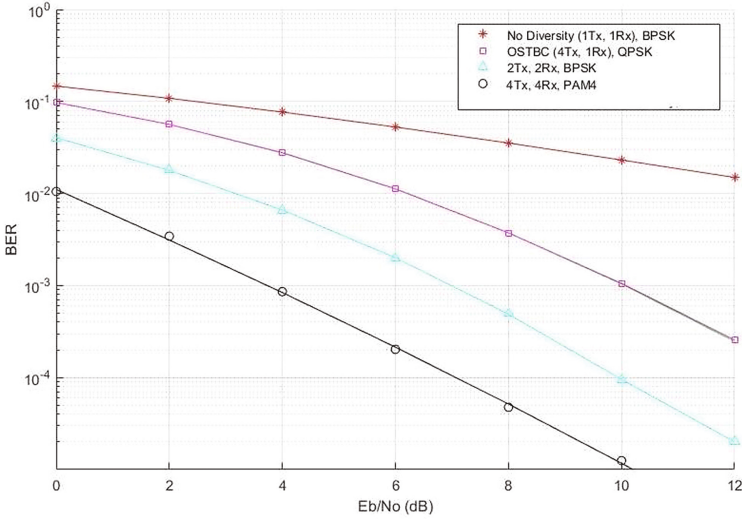


Fig. 3. BER performance of 4×4 PAM4 vs OSTBC QPSK, 1×1 and 2×2 BPSK

A comparison of BER as it varies with the symbol-energy per noise-spectral density ratio for different modulation techniques is shown in Fig. 3. It compares 4×4 PAM4 with popular 4×1 orthogonal space-time block coding (OSTBC) quadrature phase-shift keying (QPSK) and 2×2 binary phase shift keying (BPSK). This plot clearly depicts that a 4×4 PAM4 is a better choice in terms of bit-error rate.

Burn-through and cross-over ranges are important performance measures in the presence of jammer. Cross-over range is the range at which signal power from target reflection and signal power from jammer equals. Cross-over range R_{co} can be calculated for a self-screening jammer as [21]-

$$R_{co} = \sqrt{\frac{P_t G \sigma B_j}{4\pi B_r L (ERP)}} \tag{8}$$

where, P_t is the peak radar transmitter power, σ is the radar RCS, B_j is operating bandwidth of jammer, G is the gain of radar transmitter antenna, B_r is radar operating bandwidth, L is atmospheric loss, ERP is effective radiated power of radar antenna. The received signal-to-jammer plus noise ratio is given by-

$$\frac{S}{J + N} = \frac{\frac{P_t G \sigma A_r \tau}{(4\pi)^2 R^4 L}}{\frac{(ERP) A_r}{4\pi R^2 B_j} + kT_0} \tag{9}$$

Now, with this equation, we can calculate the range at which radar can properly detect the presence of a target. That range is the burn-through range of the radar [21]. It is defined as the maximum distance between radar and target, below which the target can be precisely detected by a radar. In terms of jammer,

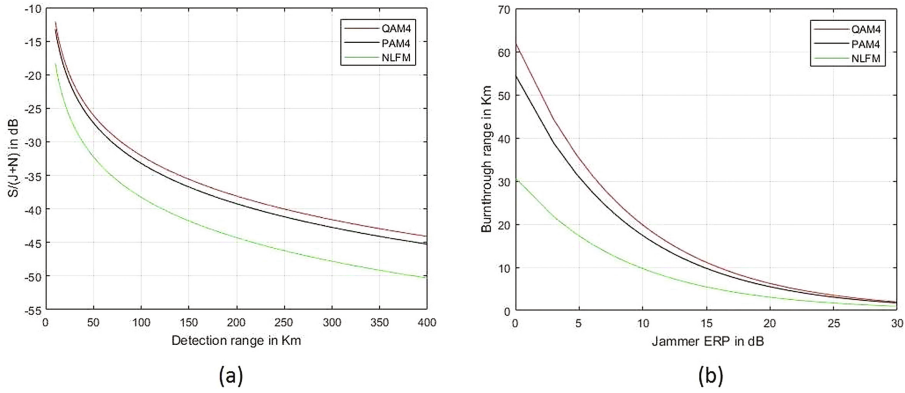


Fig. 4. (a). Detection range vs signal-to-jammer plus noise ratio, (b). Jammer ERP vs burn-through range

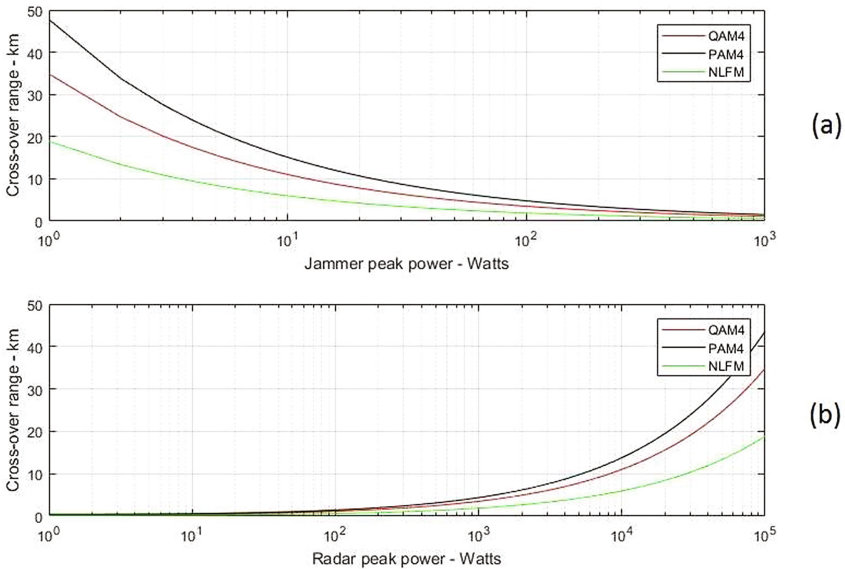


Fig. 5. (a). Jammer peak power vs cross-over range for PAM4, NLFM and QAM, (b). Radar peak power vs cross-over range for PAM4, NLFM and QAM

it is the minimum distance between radar and target beyond which target is completely masked by the jammer [15].

A comparison of detection range versus signal-to-jamming plus noise ratio for different radar signals is shown in Fig. 4, where sub-plot (b) compares jammer effective radiated power (ERP) versus the burn-through range. From Fig. 4, PAM4 demonstrates a better performance than conventional NLFM in detecting the target in the presence of a jamming signal.

The variation of cross-over range with increasing jammer power and radar transmitter power is plotted in Fig. 5; sub-plot (a) shows how increasing jammer power reduces cross-over range. It becomes evident from here that cross-over range with PAM4 is higher than with either of NLFM or QAM, for a given value of jammer power. From sub-plot (b) it is clear that PAM4 performs better for all values of transmitter power.

From the various such performance comparisons it is inferred that PAM4 performs much better than NLFM, and is very much comparable with QAM technique. The advantage of PAM over QAM is the simplicity of design of circuits for modulation and demodulation. When the aim of radar is to detect target rather than data transmission, it is preferable that the electronics and computation involved in generation and detection of signals be kept less complex. Hence, PAM4 seems to be a good choice as a radar signalling scheme.

5 Conclusions

In this paper, PAM4 is introduced as a new signaling scheme for radar systems. The effectiveness of PAM4 for radar is evaluated through metrics such as burn-through range, cross-over range and bit-error-rate. It is also compared against the NLFM technique, which is the most popular and preferred radar signaling method. Comparison is also done with QAM16, which is a popular modulation system in wireless communication. Results show that PAM4 performs better against jammers when compared to NLFM and is very well comparable with QAM16. At the same time, PAM4 needs a simple circuitry for implementation in hardware, and is also computationally less demanding during radar signal processing. It also requires lesser transmitted power compared to QAM16 and NLFM. Unlike NLFM, PAM4 can be modulated easily with digital encryption data to counter deception schemes. Hence, a thorough research and development in PAM4 for radar applications can make it a competitive technology in the future.

In the present work, the data used to modulate with PAM4 is generated as a random set of integers. A proper choice of data bits could be more powerful against jammers. Also, since MIMO model is used, a different set of orthogonal data may be transmitted through each of the MIMO transmitter element. Further, PAM4 may be considered not only in tracking radars, but also in detection and imaging radar applications such as synthetic aperture radar (SAR) as an appropriate signalling scheme.

References

1. Daniel, R.F., Paul, J.B., Muralidhar, R.: Signaling strategies for the hybrid MIMO phased-array radar. *IEEE J. Sel. Top. Sig. Process.* **4**(1), 66–78 (2010)
2. Gesbert, D., Shafi, M., Shiu, D.S., Smith, P.J., Naguib, A.: From theory to practice - an overview of MIMO space-time coded wireless systems. *IEEE J. Sel. Areas Commun.* **21**(3), 281–302 (2003)

3. Yongzhe, L., Sergiy, A.V., Aboulnasr, H.: Robust beamforming for jammers suppression in MIMO radar. In: Radar Conference. IEEE (2014)
4. Anjana, C., Shanmugha Sundaram, G.A.: Simulating the impact of wind turbine on RADAR signals in L and S band using XGtd. In: Proceedings of ICCSP 2015, pp. 199–203. IEEE (2015)
5. Li, X.-M., Luo, D., Qiu, C.-Y., Li, C.: Adaptive generalized DPCA algorithm for clutter suppression in airborne radar system. In: CIE International Conference on Radar. IEEE (2011)
6. Yunhan, D.: A new adaptive displaced phase centre antenna processor. In: International Conference on Radar. IEEE (2008)
7. Cem, S.: Digital communication jamming. Thesis, Naval Postgraduate School, California, September 2000
8. Ankarao, V., Srivatsa, S., Shanmugha Sundaram, G.A.: Evaluation of pulse compression techniques in Xband radar systems. In: International Conference on Wireless Communications Signal Processing and Networking, March 2017
9. Anritsu Corp.: High-Speed PAM Signal Generation and BER Measurements. Application Note, Anritsu (2016)
10. Yoshio, K., Hiroyasu, I., Hisato, I., Hideyuki, S.: Satellite communications using ultra wideband (UWB) signals. In: International Symposium on Advanced Radio Technologies (2004)
11. Li, X., et al.: Wireless visible light communications employing feed-forward pre-equalization and PAM-4 modulation. *J. Lightwave Technol.* **34**(8), 2049–2055 (2016)
12. Tektronix Inc.: PAM4 Signaling in High Speed Serial Technology: Test, Analysis, and Debug. Application Note, Tektronix (2015)
13. Aboulnasr, H., Sergiy, A.V.: Phased-MIMO radar: a tradeoff between phased-array and MIMO radars. *IEEE Trans. Sig. Process.* **58**(6), 3137–3151 (2010)
14. Joseph, A.J., Guy, G., Charles, B., Steven, P.L., Donald, M.C.: Configurable clutter models for radar simulations. In: Waveform Diversity and Design Conference (WDD). IEEE (2012)
15. Shlomo, F., Ben-Zion, B.: A novel approach to the burn-through range - theory and applications. In: Radar Conference. IEEE (2000)
16. Ishak, S., Ahmad, K.S., Yeoh, C.Y., Abdul, A.: Performance evaluation of distributed and co-located MIMO LTE physical layer using wireless open-access research platform. *Int. J. Electr. Energ. Electron. Commun. Eng.* **10**(10) (2016)
17. TI Inc.: RF Sampling S-band Radar Receiver, TI Designs, Texas Instruments (2016)
18. Joseph, A.D., Dale, J.M., Michael, A.E.: Variable coding and modulation experiment using NASA's space communication and navigation Testbed, NASA/TM (2016)
19. Andre, S., Andreas, A., Steffen, L.: Experimental evaluation of a (4×4) multi-mode MIMO system utilizing customized optical fusion couplers. In: Photonic Networks, 17 ITG-Symposium, Proceedings of VDE (2016)
20. Wessling, A.: Radar target modeling based on RCS measurements (2002)
21. Bessem, R.M.: Introduction to Radar Analysis. CRC Press, Boca Raton (1998)

Multi-criteria Decision Making on Lattice Ordered Multisets

V.S. Anusuya Ilamathi and J. Vimala^(✉)

Department of Mathematics, Alagappa University, Karaikudi, India
anusuyailamathimaths1@gmail.com, vimaljey@alagappauniversity.ac.in

Abstract. The crucial aspect of a multiset is the recurrence of its objects which we get it from a set and it is a new interesting mathematical notion. In this paper, lattice theory is applied to the multiset in context. We introduce a new concept for sorting things, demonstrated with lots of quantitative and qualitative attributes, and may endure in several copies. We then investigate few of the properties relevant to them. Moreover, we propose 0 and 1 as multiset depiction and identify how they play an application role in the lattice (anti-lattice) ordered multisets, where some sorting exists amid the attributes are analysed.

Keywords: Multiset · Lattices · ℓ -mset · Anti- ℓ -mset

MSC(2010): 06D72, 06B23

1 Introduction

In our practical life, due to the complexity and unreliability of decision-making problems, decision outcome may also be erratic unreliability, and so, it could be hard for decision makers to unambiguous an explicit priority relation specified which they can have only unambiguous their unreliable binary priority between choices (for details, see [2, 7, 13–15]). Various decision making instances and patterns are developed for support human to making decisions under intricate situations, but it is yet difficult to make a good decision, specifically in the intricate, dynamic, and unreliable environment.

Partially ordered data is found everywhere in our day-to-day decision-making problems due to the unreliable and dynamic environment. For instance, one alternative is better in one aspect but may be worse in another, and we used to find it hard for making a decision when multiple criteria occurred where conflicting assessments arise ever. Partial orders are more malleable than total orders to stand for incomplete, unreliable and inaccurate knowledge.

Lattice, a peculiar aspect of partial order in context. The elegant source on Lattice Theory are the books [1, 5]. Though lattice theory is built upon the notions which are simple, they can be developed to a rich network of various properties with many applications. Lattice theory has been applied to

all kinds of fields. Distributivity of lattice ordered group was developed by Natarajan and Vimala [9,10] in detail. Sorting and ordering objects by their features are the prevalent complication of decision making. Ordinal ranks are supposed to be sorted from the best to worst. Attributes may have various relative eminence. The attribute index based on the purpose of decision analysis.

Multiset theory is one of the mathematical tool to handle unreliabilities. A multiset is a collection of objects in which the repetitions of elements is significant. Relations and functions in the multiset context was established in [3] followed by [4]. For detail studies of multisets one can refer [6,8,11,12,16,17]. While handling a collection of employees' experience in a company, we need to handle entries bearing repetitions. In those situations the classical definition of set proves inadequate for the situation conferred. Here we say well-experienced employees of the company, it may mean those employees who working over thirty five years and high experience employees, it may mean those employees who working over thirty to thirty five years. In such a case, there is an order among them like fresher, low, medium, high, well experiencers. With this motivation in mind, the notion of ℓ -mset (anti ℓ -mset) arisen.

The rest of this paper is structured as: Sect. 2, provides a revision about the basic definitions based on multisets and notions for further study. Section 3 presents, the definition of ℓ -mset some of its algebraic properties. In Sect. 4, a review of the applications in which the ℓ -mset has been applied is showed. Moreover, comparative analysis is presented by using the same example as well. We conclude this paper with Sect. 5.

2 Preliminaries

In this section, some necessary definitions of multisets and lattices, existing representation of multisets, arithmetic operations between multisets are reviewed.

Definition 2.1 [6]. *Let X be any set. A multiset M drawn from X is represented by a function count M or C_M defined as $C_M: X \rightarrow \mathbb{N}$ where \mathbb{N} represents the set of all non-negative integers.*

For each $x \in X$, $C_M(x)$ is the characteristic value of x in M and indicates the number of occurrences of the element x in M . A multiset M is a set if $C_M(x) = 0$ or $1 \forall x \in X$.

The word multiset often shortened to 'mset' abbreviates the term 'multiple membership set'.

Definition 2.2 [6]. *Let M_1 and M_2 be two multisets drawn from a set X . M_1 is a sub multiset of M_2 ($M_1 \subseteq M_2$) if $C_{M_1}(x) \leq C_{M_2}(x)$ for all $x \in X$. M_1 is a proper sub multiset of M_2 ($M_1 \subset M_2$) if $C_{M_1}(x) \leq C_{M_2}(x)$ for all $x \in X$ and there exist at least one $x \in X$ such that $C_{M_1}(x) < C_{M_2}(x)$.*

Definition 2.3 [6]. *Two multisets M_1 and M_2 are equal ($M_1 = M_2$) if $M_1 \subseteq M_2$ and $M_1 \supseteq M_2$.*

Definition 2.4 [6]. *An multiset M is empty if $C_M(x) = 0$ for all $x \in X$.*

Definition 2.5 [6]. *The union of two multisets M_1 and M_2 drawn from a set X is an multiset M denoted by $M = M_1 \cup M_2$ such that $\forall x \in X, C_M(x) = \max\{C_{M_1}(x), C_{M_2}(x)\}$.*

Definition 2.6 [6]. *The intersection of two multisets M_1 and M_2 drawn from a set X is an multiset M denoted by $M = M_1 \cap M_2$ such that $\forall x \in X, C_M(x) = \min\{C_{M_1}(x), C_{M_2}(x)\}$.*

Definition 2.7 [6]. *Addition of two msets M_1 and M_2 drawn from a set X results in a new mset $M = M_1 \oplus M_2$ such that $\forall x \in X, C_M(x) = C_{M_1}(x) + C_{M_2}(x)$.*

Definition 2.8 [6]. *Addition of two msets M_1 and M_2 drawn from a set X results in a new mset $M = M_1 \ominus M_2$ such that $C_M(x) = \max\{C_{M_1}(x) - C_{M_2}(x), 0\}$.*

Definition 2.9 [6]. *Multiplication of two msets M_1 and M_2 drawn from a set X results in a new mset $M = M_1 \otimes M_2$ such that $\forall x \in X, C_M(x) = C_{M_1}(x).C_{M_2}(x)$.*

Notation [4]. Let M be an mset from X and let x appear n times in M . We denote it by $x \in^n M$. $M = \{k_1/x_1, k_2/x_2, \dots, k_n/x_n\}$ also means that M is an mset with x_1 appearing k_1 times, x_2 appearing k_2 times and so on. $[M]_x$ denotes the element x belonging to the mset M and $|[M]_x|$ denotes the cardinality of an element x in M .

Definition 2.10 [4]. *A subset C of (M, \leq) is called a chain in a partially ordered mset if every distinct pair of points from C is comparable in (M, \leq) . i.e., for all distinct $m/x, n/y$ in C , then $m/x \preceq n/y$ in (M, \leq) .*

Definition 2.11 [5]. *A poset (L, \mathfrak{R}) is a nonempty set together with a binary relation \mathfrak{R}' that satisfy the following*

- $x \mathfrak{R} x$ for all $x \in L$ (reflexivity)
- $x \mathfrak{R} y$ and $y \mathfrak{R} x \Rightarrow x = y$ for all $x, y \in L$ (anti-symmetricity)
- $x \mathfrak{R} y$ and $y \mathfrak{R} z \Rightarrow x \mathfrak{R} z$ for all $x, y, z \in L$ (transitivity).

Definition 2.12 [5]. *Let (L, \perp, \top) be a lattice. A partial ordering relation \mathfrak{R} is defined on L by $x \mathfrak{R} y$ if and only if $x \perp y = x$ and $x \top y = y$.*

Definition 2.13 [5]. *Let \mathfrak{R} be a relation defined on a set L . Then the converse relation of \mathfrak{R} denoted by \mathfrak{R}' is defined by $x \mathfrak{R} y \Leftrightarrow y \mathfrak{R}' x, x, y \in L$. If a set L forms a poset under a relation \mathfrak{R} then L forms a poset under \mathfrak{R}' . If (L, \mathfrak{R}) be a poset then the poset (L', \mathfrak{R}') , where $L' = L$ and \mathfrak{R}' is converse of \mathfrak{R} is called dual of L .*

Remark 2.14. *If \mathfrak{R} is a partial ordering, then its converse \mathfrak{R}' , is also a partial ordering. So (L, \mathfrak{R}') is also a lattice if (L, \mathfrak{R}) is a lattice. The Hasse diagram of (L, \mathfrak{R}') is obtained by turning the Hasse diagram of (L, \mathfrak{R}) upside down. Also, $x \perp y$ in (L, \mathfrak{R}) is $x \top y$ in (L, \mathfrak{R}') . A similar property holds good for $x \top y$.*

From the above remarks, we see that each lattice (L, \mathfrak{R}) determines another lattice (L, \mathfrak{R}') . (L, \mathfrak{R}') can be called the dual of (L, \mathfrak{R}) . According to the principal of duality, any valid statement about a lattice (L, \mathfrak{R}) involving $\perp, \top, \mathfrak{R}$ and \mathfrak{R}' remains valid if \perp and \top are interchanged and \mathfrak{R} and \mathfrak{R}' are interchanged.

3 On Lattice Ordered Multisets

Throughout this work, X refers a root set of a multiset and also signifies to the lattice.

Definition 3.1. A multiset M is called lattice (anti-lattice) ordered multiset whenever for the function $C_M: X \rightarrow \mathbb{N}$, $x \leq y$ implies $C_M(x) \leq C_M(y)$ [$C_M(x) \geq C_M(y)$] for all $x, y \in X$.

The word ‘lattice ordered multiset’ often shortened to ‘ ℓ -mset’ that abbreviates the term partially ordered multiple membership set in which each two-element submultiset has an infimum and a supremum.

Definition 3.2. For every $C_M(x), C_M(y) \in \mathbb{N}$, define $C_M(x) \vee C_M(y) = \max\{C_M(x), C_M(y)\}$ and $C_M(x) \wedge C_M(y) = \min\{C_M(x), C_M(y)\}$.

Example 3.3. Let $M = \{m/c, m+i/a, m+i+j/b, m+i/d, m+i+k/e\}$ with $i < j < k < m$ and $i, j, k, m \in \mathbb{N}$ be a multiset of gems annexed within a necklace. Here a, b, c, d, e depict platinum, diamond, coral, emerald, sapphire respectively. Necklace is highly designed by sapphire. Sorting of those elements is as delineated in Fig. 1 and whose tabular form is as given in below Table 1.

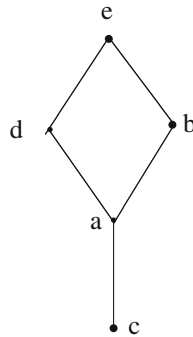


Fig. 1.

Table 1.

C_M	m	$m + i$	$m + i + j$	$m + i + k$
a	0	1	0	0
b	0	0	1	0
c	1	0	0	0
d	0	1	0	0
e	0	0	0	1

Theorem 3.4. *Let M be a ℓ -mset drawn from X , then the following are equivalent*

- (i) $x \leq y$ implies $C_M(x) \leq C_M(y) \forall x, y \in X$.
 - (ii) $C_M(x) \wedge C_M(y) = C_M(y) \forall x, y \in X$
 - (iii) $C_M(x) \vee C_M(y) = C_M(x) \forall x, y \in X, C_M(x), C_M(y) \in \mathbb{N}$.
- (i.e.,) \leq is a partially ordered relation

Proof. Follows from Definitions 3.1 and 3.2.

Theorem 3.5. *Let M be a ℓ -mset (anti ℓ -mset) drawn from X . Then*

- (i) for all $x, y \in X$, and for $C_M(x), C_M(y) \in \mathbb{N}, C_M(x \vee y)$ and $C_M(x \wedge y)$ are the least upper bound and greatest lower bound of $C_M(x)$ and $C_M(y)$ respectively.
- (ii) M is a partially ordered set.

Proof. (i) Suppose $C_M(x \vee y)$ is not a least upper bound of $C_M(x)$ and $C_M(y)$. Then there is $a \in X$ such that $x \leq a \leq x \vee y$ and $y \leq a \leq x \vee y$. Then $x \vee y \leq a \vee a \leq x \vee y \Rightarrow C_M(x \vee y) \leq C_M(a \vee a) \leq C_M(x \vee y) \Rightarrow C_M(a) = C_M(x \vee y)$ which is a contradiction. Hence $C_M(x \vee y)$ is the least upper bound of $C_M(x)$ and $C_M(y)$. The proof for greatest lower bound will be made by an analogous way.

(ii) By (i) and Example 3.3, M is a partially ordered set.

Theorem 3.6. *Let M be a ℓ -mset (anti ℓ -mset) drawn from X . Then for every $C_M(x), C_M(y) \in \mathbb{N}$, the following are hold*

Idempotent

$$C_M(x) \vee C_M(x) = C_M(x) \text{ and } C_M(x) \wedge C_M(x) = C_M(x) \forall x, y \in X$$

Commutative

$$C_M(x) \vee C_M(y) = C_M(y) \vee C_M(x) \text{ and } C_M(x) \wedge C_M(y) = C_M(y) \wedge C_M(x) \forall x, y \in X$$

Associative

$$(C_M(x) \vee C_M(y)) \vee C_M(z) = C_M(x) \vee (C_M(y) \vee C_M(z)) \text{ and } (C_M(x) \wedge C_M(y)) \wedge C_M(z) = C_M(x) \wedge (C_M(y) \wedge C_M(z)) \forall x, y \in X$$

Absorption

$$C_M(x) \vee (C_M(x) \wedge C_M(y)) = C_M(x) \text{ and } C_M(x) \wedge (C_M(x) \vee C_M(y)) = C_M(x) \text{ for all } x, y \in X$$

Proof. It is straightforward.

Theorem 3.7. *Let M_1 and M_2 be two ℓ -msets drawn from X_1 and X_2 respectively. Then $M_1 \cap M_2$ is also a ℓ -mset.*

Proof. This theorem is trivial for $M_1 \cap M_2 = \emptyset$. Suppose $M_1 \cap M_2$ is non-empty, then both M_1 and M_2 inherit the partial order by any attribute. Hence for every $a_1 \leq_{X_1} a_2$ we have $C_{M_1}(a_1) \leq C_{M_1}(a_2)$ for $a_1, a_2 \in X_1$ and for

$C_{M_1}(a_1), C_{M_1}(a_2) \in \mathbb{N}$. Also, for every $b_1 \leq_{X_2} b_2$ we have $C_{M_2}(b_1) \leq C_{M_2}(b_2)$ for $b_1, b_2 \in X_2$ and for $C_{M_2}(b_1), C_{M_2}(b_2) \in \mathbb{N}$. Therefore for every $s_1, s_2 \in X_3, C_{M_1 \cap M_2}(s_1) = \min\{C_{M_1}(s_1), C_{M_2}(s_1)\} \leq \min\{C_{M_1}(s_2), C_{M_2}(s_2)\} = C_{M_1 \cap M_2}(s_2)$ for $s_1 \leq_{X_3} s_2$ where X_3 is the root set of $M_1 \cap M_2$. Thus $M_1 \cap M_2$ is ℓ -mset. Such an analogy made for anti- ℓ -mset.

Theorem 3.8. *Let M_1 and M_2 be two ℓ -mssets drawn from X_1 and X_2 respectively. Then $M_1 \cup M_2$ is also ℓ -mset.*

Proof. Given M_1 and M_2 are two ℓ -mssets drawn from X_1 and X_2 respectively and so, they inherit the partial order Thus for any $a_1 \leq_{X_1} a_2$ we have $C_{M_1}(a_1) \leq C_{M_1}(a_2) \forall a_1, a_2 \in X_1$ and for $C_{M_1}(a_1), C_{M_1}(a_2) \in \mathbb{N}$. Also, for every $b_1 \leq_{X_2} b_2$ we have $C_{M_2}(b_1) \leq C_{M_2}(b_2)$ for $b_1, b_2 \in X_2$ and for $C_{M_2}(b_1), C_{M_2}(b_2) \in \mathbb{N}$. Thus for every $s_1, s_2 \in X_3, C_{M_1 \cup M_2}(s_1) = \max\{C_{M_1}(s_1), C_{M_2}(s_1)\} \leq \max\{C_{M_1}(s_2), C_{M_2}(s_2)\} = C_{M_1 \cup M_2}(s_2)$ for $s_1 \leq_{X_3} s_2$ where X_3 is the root set of $M_1 \cup M_2$. Hence $M_1 \cup M_2$ is a ℓ -mset. The result can be shown for anti ℓ -mset by the similar manner.

Theorem 3.9. *Let M be a ℓ -mset (anti ℓ -mset) drawn from X . Then $M \wedge M = \{x \wedge y : x, y \in X\}$ is a ℓ -mset.*

Proof. Since M is a ℓ -mset, for every $x_1 \leq_X y_1$ we have $C_M(x_1) \leq C_M(y_1)$ for $x_1, y_1 \in X$. Also, for every $x_2 \leq_X y_2$ we have $C_M(x_2) \leq C_M(y_2)$ for $x_2, y_2 \in X$. Hence for every $x_1, x_2, y_1, y_2 \in X, C_{M \wedge M}(x_1 \wedge x_2) = C_M(x_1) \wedge C_M(x_2) = \min\{C_M(x_1), C_M(x_2)\} \leq \min\{C_M(y_1), C_M(y_2)\} = C_M(y_1) \wedge C_M(y_2) = C_{M \wedge M}(y_1 \wedge y_2)$ for $x_1 \wedge x_2 \leq_X y_1 \wedge y_2$.

Theorem 3.10. *If A and B are ℓ -mssets (anti ℓ -mssets), then $A \wedge B = \{x \wedge y : x \in A, y \in B\}$ is also ℓ -mset (anti ℓ -mset).*

Proof. Since A and B are ℓ -mssets, for every $x_1 \leq_X x_2$ we have $C_A(x_1) \leq C_A(x_2)$ for $x_1, x_2 \in X$. Also, for every $y_1 \leq_Y y_2$ we have $C_B(y_1) \leq C_B(y_2)$ for $y_1, y_2 \in Y$. Hence for every $x_1, x_2 \in X, y_1, y_2 \in Y, C_{A \wedge B}(x_1 \wedge y_1) = C_A(x_1) \wedge C_B(y_1) = \min\{C_A(x_1), C_B(y_1)\} \leq \min\{C_A(x_2), C_B(y_2)\} = C_A(x_2) \wedge C_B(y_2) = C_{A \wedge B}(x_2 \wedge y_2)$ for $x_1 \wedge y_1 \leq_Z x_2 \wedge y_2$.

Theorem 3.11. *Let M_1 and M_2 be two ℓ -mssets (anti ℓ -mssets) drawn from X_1 and X_2 respectively. Then $M_1 \oplus M_2$ is also a ℓ -mset.*

Proof. Since $C_{M_1}(x) + C_{M_2}(x) \leq C_{M_1}(y) + C_{M_2}(y)$, thus we get the result.

Theorem 3.12. *If A and B are two ℓ -mssets (anti ℓ -mssets), then $A \otimes B$ is also ℓ -mset (anti ℓ -mset).*

Proof. Since A and B are two ℓ -mssets drawn from X , we have $C_A(x).C_B(x) \leq C_A(y).C_B(y)$ for every $x \leq_X y$, thereby we get the result.

Theorem 3.13. *If A is a ℓ -mset and B is an anti- ℓ -mset, then $A \ominus B$ is also a ℓ -mset.*

Proof. Since A is ℓ -mset, for each $a_1 \leq_{X_1} a_2$ we have $C_A(a_1) \leq C_A(a_2)$ for $a_1, a_2 \in X_1$. Since B is an anti- ℓ -mset, for every $b_1 \leq_{X_2} b_2$ we have $C_B(b_1) \geq C_B(b_2)$ for $b_1, b_2 \in X_2$ and so, for each $a \leq_{X_3} b, C_{A \oplus B}(a) = \max\{C_A(a) - C_B(a), 0\} \leq \max\{C_A(b) - C_B(b), 0\} = C_{A \oplus B}(b)$ for $a, b \in X_3$.

Theorem 3.14. *Every chain is an ℓ -mset.*

Proof. By definition of chain, for every distinct x and y in M , we have $C_M(x)/x \leq C_M(y)/y$ in (M, \leq) . i.e., $x \leq y$ and $C_M(x) \leq C_M(y) \forall x, y \in M$. Hence M is an ℓ -mset.

4 Application

The theoretical model based on multisets is used to study the structure of multicriterial alternatives. The basic concepts of multiset theory is considered. In many cases, it is difficult for decision-makers to precisely express a preference when attempting to solve multi-criteria decision-making (MCDM) problems with inaccurate, uncertain or incomplete information.

Example 4.1. *Consider a problem for retail shop keeper to buy the most appropriate invest case (count) for products. Ofcourse, buying in bulk can sometimes save his money, but those giant packages of everyday items are not always the best deal. It is important to compare the per-unit prices because he'll often find that smaller packages cost less per unit than supersize containers. So, the shop collect the customers feed back about the products. Then the products sort out according to grades as $p_1 \leq p_2 \leq p_3 \leq p_4 \leq p_5 \leq p_6$ where the product p_1 belonging into the attribute low quality, p_2 in medium quality, p_3 in satisfactory quality, p_4 in good quality, p_5 in high quality, p_6 in very high quality. This order is made by customers feedback. Here count of the product is the sales of product which are calculated by shop accounts department in the time of every four months. Product and their sales form ℓ -msets which is shown in the following Tables 2, 3 and 4.*

Table 2.

C_{M_1}	n_1	n_2	n_3	n_4	n_5
p_1	1	0	0	0	0
p_2	1	0	0	0	0
p_3	0	1	0	0	0
p_4	0	1	0	0	0
p_5	0	1	0	0	0
p_6	0	0	1	0	0

Table 3.

C_{M_3}	n_1	n_2	n_3	n_4	n_5
p_1	1	0	0	0	0
p_2	0	1	0	0	0
p_3	0	1	0	0	0
p_4	0	1	0	0	0
p_5	0	0	0	1	0
p_6	0	0	0	1	0

Table 4.

C_{M_4}	n_1	n_2	n_3	n_4	n_5
p_1	1	0	0	0	0
p_2	0	1	0	0	0
p_3	0	1	0	0	0
p_4	0	1	0	0	0
p_5	0	0	1	0	0
p_6	0	0	1	0	0

Table 5.

C_{M_4}	n_1	n_2	n_3	n_4	n_5
p_1	1	0	0	0	0
p_2	0	0	1	0	0
p_3	0	0	1	0	0
p_4	0	0	0	1	0
p_5	0	0	0	0	1
p_6	0	0	0	0	1

In the above table, there is a $k \in \mathbb{N}$ such that cases (counts) n_1 contains k number of products; n_2 contains $2k$ number of products; n_3 contains $3k$ number of products and so on. If k number of products sold, then write 1 in n_1 and 0 for others; If $2k$ number of products sold, we write 1 in n_2 and 0 for others, and so on. i.e., $C_M(p_i) = n_j$ for some j . From this, we can see $n_1 < n_2 < n_3 < n_4 < n_5$.

Since M_1, M_2, M_3, M_4 are lattice ordered finite multisets, their union forms a lattice ordered multiset which is shown in Table 5. i.e., If $p_{i,l}, p_{i,h}$ stand for lowest and highest sale of p_i respectively, then $p_{1,l} = n_1, p_{2,l} = n_1, p_{3,l} = n_2, p_{4,l} = n_2, p_{5,l} = n_2, p_{6,l} = n_3$ and $p_{1,h} = n_2, p_{2,h} = n_3, p_{3,h} = n_3, p_{4,h} = n_4, p_{5,h} = n_5, p_{6,h} = n_5$ we have $C_M(p_1) \leq C_M(p_2) \leq C_M(p_3) \leq C_M(p_4) \leq C_M(p_5) \leq C_M(p_6)$ Also we have Table 6 as a ℓ -mset by applying $M \wedge M$ is a ℓ -mset (Table 7).

Table 6.

C_M	n_1	n_2	n_3	n_4	n_5
p_1	1	1	0	0	0
p_2	1	1	1	0	0
p_3	0	1	1	0	0
p_4	0	1	1	1	0
p_5	0	1	1	1	1
p_6	0	0	1	1	1

Table 7.

$C_{M \wedge M}$	n_1	n_2	n_3	n_4	n_5
$p_1 \wedge p_1$	1	1	0	0	0
$p_1 \wedge p_2$	1	1	0	0	0
$p_1 \wedge p_3$	0	1	0	0	0
$p_1 \wedge p_4$	0	1	0	0	0
$p_1 \wedge p_5$	0	1	0	0	0
$p_1 \wedge p_6$	0	0	0	0	0
$p_2 \wedge p_2$	1	1	1	0	0
$p_2 \wedge p_3$	0	1	1	0	0
$p_2 \wedge p_4$	0	1	1	0	0
$p_2 \wedge p_5$	0	1	1	0	0
$p_2 \wedge p_6$	0	1	0	0	0
$p_3 \wedge p_3$	0	1	1	0	0
$p_3 \wedge p_4$	0	1	1	0	0
$p_3 \wedge p_5$	0	1	1	0	0
$p_3 \wedge p_6$	0	0	1	0	0
$p_4 \wedge p_4$	0	1	1	1	0
$p_4 \wedge p_5$	0	1	1	1	0
$p_4 \wedge p_6$	0	0	1	1	0
$p_5 \wedge p_5$	0	1	1	1	1
$p_5 \wedge p_6$	0	0	1	1	1
$p_6 \wedge p_6$	0	0	1	1	1

Table 8.

	n_1	n_2	n_3	n_4	n_5
Score	3	16	14	6	3

From the above Table 8, we have $n_1 = n_5 < n_4 < n_3 < n_2$ and n_2 is the best choice. Verified the proposed approach by SAW. Simply Additive Weighting (SAW) method is probably the best known and most widely used MADM method (see [13–15]) and it is applied to verify the selection of count as seen in Example 4.1.

Algorithm is presented below:

1. Let $A = (a_1, a_2, a_3, \dots, a_m)$ be a set of alternatives and let $C = (c_1, c_2, c_3, \dots, c_n)$ be a set of criteria
2. Construct the decision matrix in terms of d_{ij} , where d_{ij} is the rating of alternative A_i with respect to criterion C_i .
3. Construct the normalized decision matrix $r_{ij} = d_{ij}/\max(d_{ij})$.
4. Construct the weighted normalized decision matrix $v_{ij} = w_j r_{ij}$, $\sum_{j=1}^m w_j = 1$
5. Calculate the score of each alternative: $S_i = \sum_{j=1}^m v_{ij}$, $i = 1, 2, \dots, n$
6. Select the best alternative: Best alternative = $\max_{1 \leq i \leq n} S_i$

The following is the decision matrix for five alternatives and six products (ordered by attributes) is written here.

$$\begin{matrix}
 & p_1 & p_2 & p_3 & p_4 & p_5 & p_6 \\
 n_1 & \left(\begin{matrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{matrix} \right) \\
 n_2 \\
 n_3 \\
 n_4 \\
 n_5
 \end{matrix}$$

Next we have to find the weight of each attribute by the following matrix

$$\begin{matrix}
 & p_1 & p_2 & p_3 & p_4 & p_5 & p_6 \\
 n_1 & \left(\begin{matrix} \frac{1}{2} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & 0 \\ 0 & \frac{1}{3} & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{3} \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{4} & \frac{1}{3} \\ 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{3} \end{matrix} \right) \\
 n_2 \\
 n_3 \\
 n_4 \\
 n_5
 \end{matrix}$$

$w_1 = 0.14, w_2 = 0.32, w_3 = 0.29, w_4 = 0.15, w_5 = 0.1$ and $\sum_i w_i = 1$.

Next $v_{ij} = w_j r_{ij}$ where $r_{ij} = \frac{d_{ij}}{\max(d_{ij})}$

Below matrix represents r_{ij} from matrix II

$$\begin{matrix}
 & p_1 & p_2 & p_3 & p_4 & p_5 & p_6 \\
 n_1 & \left(\begin{matrix} 1 & 0.6 & 0 & 0 & 0 & 0 \\ 1 & 0.6 & 1 & 0.6 & 0.5 & 0 \\ 0 & 0.6 & 1 & 0.6 & 0.5 & 0.6 \\ 0 & 0 & 0 & 1 & 0.75 & 1 \\ 0 & 0 & 0 & 0 & 0.75 & 1 \end{matrix} \right) \\
 n_2 \\
 n_3 \\
 n_4 \\
 n_5
 \end{matrix}$$

v_{ij} shown in following matrix

$$\begin{matrix} & p_1 & p_2 & p_3 & p_4 & p_5 & p_6 \\ \begin{matrix} n_1 \\ n_2 \\ n_3 \\ n_4 \\ n_5 \end{matrix} & \begin{pmatrix} 0.14 & 0.084 & 0 & 0 & 0 & 0 \\ 0.32 & 0.192 & 0.32 & 0.192 & 0.16 & 0 \\ 0 & 0.174 & 0.29 & 0.174 & 0.145 & 0.174 \\ 0 & 0 & 0 & 0.15 & 0.113 & 0.15 \\ 0 & 0 & 0 & 0 & 0.075 & 0.1 \end{pmatrix} \end{matrix}$$

Next $S_i = \sum_{j=1}^5 v_{ij}, i = 1, 2, \dots, 6$ we have

Table 9.

Count	S_i	Rank
n_1	0.224	3rd
n_2	1.184	1st
n_3	0.957	2nd
n_4	0.413	4th
n_5	0.175	5th

From both the approaches, we can get maximum priority case as n_2 . Such a case will be profitable and we can avoid abundance of goods in warehouses and save money when buying such cases (Table 9).

Example 4.2. *This is an example for score evaluation. Consider a problem for analysing the performance of the university and to increasing the knowledge of the students. Thus making them ease for job seeking and as well as to excel them in irrespective of all the fields. For the above, we then gather the particulars of syllabi and centum scorers. Favorably, they form an anti-l-mset. In such a case, let us consider the whole score board with the syllabi. Suppose that there are few kinds of scores: centum (m_5), high score (m_4), below average score (m_1), average score (m_2), good score (m_3) with the order $m_1 \leq m_2 \leq m_3 \leq m_4 \leq m_5$ are under evaluation according to following alternatives: repeated syllabi (s_1), old syllabus with a few changes (s_2), old syllabus with latest trends (s_3), new syllabus (s_4), new syllabus with global connects (s_5), syllabus with real world applications (s_6), advanced syllabus (s_7). The order of syllabi is shown in Fig. 2: $s_1 \leq s_4 \leq s_5 \leq s_7$ and $s_1 \leq s_2 \leq s_3 \leq s_6 \leq s_7$ (Tables 10 and 11).*

If the Institute has T number of students, then there is $k \in \mathbb{N}$ such that $k = \frac{T}{5}$ and write 1 in m_i if k number of students obtained the score m_i ; otherwise 0. i.e., If $s_{i,k,l}, s_{i,k,h}$ stand for the lowest and highest score which are obtained by at least k number of students for the syllabi s_i respectively, then $s_{1,k,l} = m_3, s_{2,k,l} = m_2, s_{3,k,l} = m_2, s_{4,k,l} = m_2, s_{5,k,l} = m_1, s_{6,k,l} = m_1, s_{7,k,l} = m_1$

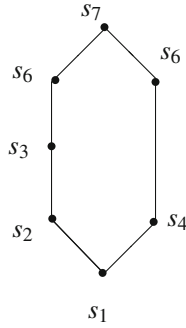


Fig. 2.

Table 10.

C_A	m_1	m_2	m_3	m_4	m_5
s_1	0	0	1	0	1
s_4	0	1	1	1	1
s_5	1	0	1	1	0
s_7	1	0	1	1	0

Table 11.

C_B	m_1	m_2	m_3	m_4	m_5
s_1	0	0	1	0	1
s_2	0	1	0	1	1
s_3	0	1	0	1	1
s_6	1	0	1	1	0
s_7	1	0	1	1	0

and $s_{1,k,h} = m_5, s_{2,k,h} = m_5, s_{3,k,h} = m_5, s_{4,k,h} = m_5, s_{5,k,h} = m_4, s_{6,k,h} = m_4, s_{7,k,h} = m_4$. In that sense, we have $C_A(s_7) \leq C_A(s_5) \leq C_A(s_4) \leq C_A(s_1)$ and $C_B(s_7) \leq C_B(s_6) \leq C_B(s_3) \leq C_B(s_2) \leq C_B(s_1)$ i.e., The above A and B are two anti-lattice ordered submultisets of anti- ℓ -mset M and as $A \wedge B$ is an anti- ℓ -mset, we have the following Table 12.

Many generalized items of a syllabus can be amplified in a specific curriculum to maximize efficient learning by clarifying student understanding of specified material such as grading policy (Table 13).

Table 12.

$C_{A \wedge B}$	m_1	m_2	m_3	m_4	m_5
$s_1 \wedge s_1$	0	0	1	0	1
$s_1 \wedge s_2$	0	0	0	0	1
$s_1 \wedge s_3$	0	0	0	0	1
$s_1 \wedge s_6$	0	0	1	0	0
$s_1 \wedge s_7$	0	0	1	0	0
$s_4 \wedge s_1$	0	0	1	0	1
$s_4 \wedge s_2$	0	1	0	1	1
$s_4 \wedge s_3$	0	1	0	1	1
$s_4 \wedge s_6$	0	0	1	1	0
$s_4 \wedge s_7$	0	0	1	1	0
$s_5 \wedge s_1$	0	0	1	0	0
$s_5 \wedge s_2$	0	0	0	1	0
$s_5 \wedge s_3$	0	0	0	1	0
$s_5 \wedge s_6$	1	0	1	1	0
$s_5 \wedge s_7$	1	0	1	1	0
$s_7 \wedge s_2$	0	0	0	1	0
$s_7 \wedge s_3$	0	0	0	1	0
$s_7 \wedge s_6$	1	0	1	1	0
$s_7 \wedge s_7$	1	0	1	1	0

Table 13.

Score	m_1	m_2	m_3	m_4	m_5
Total	4	2	11	12	6
Rank	4th	5th	2nd	1st	3rd

At any changes of syllabi, ‘high-score’ has been maintained here.

From the above table, we have m_4 as the ‘all-time’ score.

In addition, this Illustration has been verified by SAW. By applying this algorithm, we have the following is the decision matrix for five alternatives and five syllabi (ordered by attributes) is written here.

$$\begin{matrix}
 & s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 \\
 m_1 & \left(\begin{matrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 \end{matrix} \right) \\
 m_2 & \\
 m_3 & \\
 m_4 & \\
 m_5 &
 \end{matrix}$$

Next step is to find the weight of each attribute by the following matrix

$$\begin{matrix}
 & s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 \\
 m_1 & \left(\begin{matrix} 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{matrix} \right) \\
 m_2 & \left(\begin{matrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{4} & 0 & 0 & 0 \end{matrix} \right) \\
 m_3 & \left(\begin{matrix} \frac{1}{2} & 0 & \frac{1}{4} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \end{matrix} \right) \\
 m_4 & \left(\begin{matrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{4} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{matrix} \right) \\
 m_5 & \left(\begin{matrix} \frac{1}{2} & \frac{1}{3} & \frac{1}{3} & \frac{1}{4} & 0 & 0 & 0 \end{matrix} \right)
 \end{matrix}$$

$w_1 = 0.18, w_2 = 0.17, w_3 = 0.19, w_4 = 0.3, w_5 = 0.16$ and $\sum_i w_i = 1$.

Next $v_{ij} = w_j r_{ij}$ where $r_{ij} = \frac{d_{ij}}{\max(d_{ij})}$

Below matrix represents r_{ij} from matrix II

$$\begin{matrix}
 & s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 \\
 m_1 & \left(\begin{matrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{matrix} \right) \\
 m_2 & \left(\begin{matrix} 0 & 1 & 1 & 0.75 & 0 & 0 & 0 \end{matrix} \right) \\
 m_3 & \left(\begin{matrix} 1 & 0 & 0 & 0.5 & 0.6 & 0.6 & 0.6 \end{matrix} \right) \\
 m_4 & \left(\begin{matrix} 0 & 1 & 1 & 0.75 & 1 & 1 & 1 \end{matrix} \right) \\
 m_5 & \left(\begin{matrix} 1 & 0.6 & 0.6 & 0.5 & 0 & 0 & 0 \end{matrix} \right)
 \end{matrix}$$

v_{ij} shown in following matrix

$$\begin{matrix}
 & s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 \\
 m_1 & \left(\begin{matrix} 0 & 0 & 0 & 0 & 0.18 & 0.18 & 0.18 \end{matrix} \right) \\
 m_2 & \left(\begin{matrix} 0 & 0.17 & 0.17 & 0.13 & 0 & 0 & 0 \end{matrix} \right) \\
 m_3 & \left(\begin{matrix} 0.19 & 0 & 0 & 0.095 & 0.11 & 0.11 & 0.11 \end{matrix} \right) \\
 m_4 & \left(\begin{matrix} 0 & 0.3 & 0.3 & 0.225 & 0.3 & 0.3 & 0.3 \end{matrix} \right) \\
 m_5 & \left(\begin{matrix} 0.16 & 0.096 & 0.096 & 0.08 & 0 & 0 & 0 \end{matrix} \right)
 \end{matrix}$$

Next $S_i = \sum_{j=1}^5 v_{ij}, i = 1, 2, \dots, 5$ we have

Table 14.

<i>Count</i>	S_i	<i>Rank</i>
m_1	0.54	3rd
m_2	0.47	4th
m_3	0.62	2nd
m_4	1.73	1st
m_5	0.43	5th

From both the approaches, we can get m_4 as all-time score (Table 14).

5 Conclusion

Notions of lattice ordered msets and anti-lattice ordered msets have been introduced. These notions are quite handy where sorting exists amid linguistic terms. One can also verify the proposed approach by other comparative analysis such as *TOPSIS*, etc. We then discuss few properties of operations of ℓ -msets (anti ℓ -msets) which are applied in the decision making problems. To widen this work, one can survey the other properties of ℓ -msets (anti ℓ -msets).

Acknowledgement. The authors thank the editors and anonymous referees for their careful reading of the manuscript and constructive comments that led to an improved version of this paper.

Conflicts of Interest

The authors declare no conflicts of interest. The authors alone are responsible for the content of this manuscript.

References

1. Birkhoff, G.: Lattice Theory, 3rd edn. American Mathematical Society, Providence (1967)
2. Chen, S., Liu, J., Wang, H., Augusto, J.C.: Ordering based decision making-a survey. *Inf. Fusion* **14**(4), 521–531 (2013)
3. Girish, K.P., John, S.J.: Relations and functions in the multiset context. *Inf. Sci.* **179**, 758–768 (2009)
4. Girish, K.P., John, S.J.: General relations between partially ordered multisets and their chains and antichains. *Math. Commun.* **14**(2), 193–206 (2009)
5. Grätzer, G.: General Lattice Theory, 2nd edn. Birkhauser, Boston (2003)
6. Jena, S.P., Ghosh, S.K., Tripathy, B.K.: On the theory of bags and lists. *Inf. Sci.* **132**, 241–254 (2001)
7. Liu, J., Xu, Y., Ruan, D., Martinez, L.: A lattice-valued linguistic-based decision making method. In: The Proceedings of 2005 IEEE Conference on Granular Computing, Beijing, China, pp. 199–202 (2005)
8. Nazmul, S., Majumdar, P., Samanta, S.K.: On multisets and multigroups. *Ann. Fuzzy Math. Inf.* **6**(3), 643–656 (2013)
9. Natarajan, R., Vimala, J.: Distributive l-ideal in commutative Lattice ordered group. *Acta Cienc. Indica* **33**(2), 517 (2007)
10. Natarajan, R., Vimala, J.: Distributive convex l- subgroup. *Acta Ciencia Indica* **XXXIII**(4), 1795 (2007)
11. Singh, D., Ibrahim, A.M., et al.: An overview of the applications of multisets. *Novi Sad J. Math.* **37**(2), 73–92 (2007)
12. Syropoulos, A.: Mathematics of multisets. In: Calude, C., Paun, G., Rozenberg, G., Salomaa, A. (eds.) *Multiset Processing. Lecture Notes in Computer Science*, vol. 2235, pp. 347–358. Springer, Heidelberg (2001)
13. Triantaphyllou, E.: *Multi-Criteria Decision Making: A Comparative Study*. Kluwer Academic Publishers, Dordrecht (2000). p. 320. ISBN 0-7923-6607-7. (now Springer)
14. Chen, T.Y.: Comparative analysis of SAW and TOPSIS based on interval valued fuzzy sets: discussions on score functions and weight constraints. *Expert Syst. Appl.* **39**, 1848–1861 (2012)

15. Gayatri, V., Misal Chetan, S.: Comparative study of different multi criteria decision making methods. *Int. J. Adv. Comput. Theory Eng.* **2**(4), 2319–2526 (2013)
16. Blizard, W.D.: Multiset theory. *Notre Dame J. Form. Log.* **30**(1), 36–66 (1989)
17. Yager, R.R.: On the theory of bags. *Int. J. Gen. Syst.* **13**(1), 23–37 (1986)

Author Index

A

Abhishek, S.N., 370
Abraham, Kevin Thomas, 317
Adigun, Matthew, 28
Agarwal, Utkarsh, 232
Alcantud, José Carlos R., 149
Amudha, J., 345
Anoop, V.S., 123
Anusuya Ilamathi, V.S., 401
Aravinth, A., 370
Arunkumar, B., 102
Asharaf, S., 123
Ashoor, Tharic, 317
Ashwin, Manikandan, 317
Attar, Vahida, 174

B

Balaji, Ashwin, 66
Balakrishnan, P., 14
Bandari, Sabitha, 41
Bansal, Saumya, 163
Bedi, Punam, 163
Bhatia, Deepika, 163
Bongir, Amit, 174

C

Chaudhary, Shubham, 232
Christopher, B., 380

D

Dara, Suresh, 41
Deepak, P., 123
Deepika, M.G., 243
Devi, Salam Jayachitra, 1
Dhakad, Gourav, 189
Dhawan, Ayushi, 243
Divya, M., 370

F

Fatima, Kaleem, 74
Fauzia, Salma, 74

G

Gautam, Anjali, 163
Gharahbagh, A. Alavi, 298
Ghatol, Ashok, 84
Gonge, Sudhanshu Suhas, 84
Gopal, Madhav, 255
Gunpath, R.P., 136
Gupta, Deepa, 327
Gupta, Tanmay, 232

H

Hajihashemi, V., 298
Harilal, O.P., 222
Hrudya, P., 222

I

Indhu, G., 308

J

Janardhanan, Ramanand, 174
Jannu, Srikanth, 41
Jeyakumar, Gurusamy, 317
Jha, Girish Nath, 255

K

K., Harikumar, 268
K.E., Karthick, 66
K.P., Soman, 268
K.V., Vineetha, 204
Keerthana, N.K., 370
Kiruthika, S., 380
Kousalya, G., 102
Kumar, Katha Kishor, 41
Kurup, Dhanesh G., 204

L

Lakshmi, S., 380

M

M, Supriya, 112

M., Sabdhi, 268

Mathew, Terry Jacob, 149

Menon, Hema P., 356

Menon, Vijay Krishna, 268

Mudali, Pragasen, 28

Mugunth Krishnan, R., 380

Murugaraj, Bharathi, 345

N

Nair, Priyanka C., 327

Nayar, Ravi C., 327

Nitheesh, A.S., 356

O

Oki, Olukayode A., 28

Olwal, Thomas O., 28

P

Pandey, Shekhar, 112

Pati, Jayadeep, 189

Prabakaran, Poornachandaran, 222

Prakash, Parvathy, 54

Praveen, K., 308

Priyanka, Rajan, 370

Pudaruth, S., 136

R

Raj, Pethuru, 14

Ram, Amritanshu, 327

Ramar, K., 54

Rao, Raghavendra, 327

S

S., Saravanan, 66

Sajith, Anand, 66

Sethumadhavan, M., 308

Shaji, Ameena, 54

Shanmugha Sundaram, G.A., 380, 390

Shastri, Sunil Suresh, 327

Sherly, Elizabeth, 149

Shrivastava, Abhilash, 112

Shukla, K.K., 189

Singh, Buddha, 1

Soman, K.P., 212

Soyjaudah, K.M.S., 136

Srivatsa, S., 390

Subbulakshmi, S., 54

Sujadevi, V.G., 212

Sundar, Darshak, 317

Surendran, K., 222

Swarnkar, Krishnkant, 189

T

Thampi, Sabu M., 279

V

Vasudevan, Shiram K., 370

Venkatesh, Veeramuthu, 14

Vidyadharan, Divya S., 279

Vimala, J., 401

Vinayakumar, R., 212

Y

Yadav, Jyoti, 232