ADVANCES IN INTELLIGENT SYSTEMS AND COMPUTING 177

Natarajan Meghanathan
Dhinaharan Nagamalai
Nabendu Chaki (Eds.)

Advances in Computing and Information Technology

Springer

Advances in Intelligent Systems and Computing 177

Editor-in-Chief

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

For further volumes:
http://www.springer.com/series/11156

Natarajan Meghanathan, Dhinaharan Nagamalai,
and Nabendu Chaki (Eds.)

Advances in Computing and Information Technology

Proceedings of the Second International Conference on Advances in Computing and Information Technology (ACITY) July 13–15, 2012, Chennai, India – Volume 2

Springer

Editors
Dr. Natarajan Meghanathan
Department of Computer Science
Jackson State University
Jackson
USA

Dr. Dhinaharan Nagamalai
Wireilla Net Solutions PTY Ltd
Melbourne
VIC
Australia

Dr. Nabendu Chaki
Department of Computer Science &
Engineering
University of Calcutta
Calcutta
India

ISSN 2194-5357
ISBN 978-3-642-31551-0
DOI 10.1007/978-3-642-31552-7
Springer Heidelberg New York Dordrecht London

e-ISSN 2194-5365
e-ISBN 978-3-642-31552-7

Library of Congress Control Number: 2012940793

Preface

The Second International Conference on Advances in Computing and Information Technology (ACITY-2012) was held in Chennai, India, during July 13–15, 2012. ACITY attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West. The goal of this conference series is to bring together researchers and practitioners from academia and industry and share cutting-edge development in the field. The conference will provide an excellent international forum for sharing knowledge and results in theory, methodology and applications of Computer Science and Information Technology. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of Computer Science and Information Technology.

The ACITY-2012 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the conference. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer-review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically. The overall acceptance rate of ACITY-2012 is less than 20%. Extended versions of selected papers from the conference will be invited for publication in several international journals. All these efforts undertaken by the Organizing and Technical Committees led to an exciting, rich and a high quality technical conference program, which featured high-impact presentations for all attendees to enjoy, appreciate and expand their expertise in the latest developments in various research areas of Computer Science and Information Technology. In closing, ACITY-2012 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. We would like to thank the General and Program Chairs, organization staff, the members of the Technical

Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research.

It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

Natarajan Meghanathan
Dhinaharan Nagamalai
Nabendu Chaki

Organization

General Chairs

David C. Wyld	Southeastern Louisiana University, USA
E.V. Krishnamurthy	Australian National University, Australia
Jae Kwang Lee	Hannam University, South Korea
Jan Zizka	SoNet/DI, FBE, Mendel University in Brno, Czech Republic
V.L. Narasimhan	Pentagram R&D Intl. Inc., New Bern, USA
Michal Wozniak	Wroclaw University of Technology, Poland

Steering Committee

Abdul Kadhir Ozcan	Karatay University, Turkey
Brajesh Kumar Kaushik	Indian Institute of Technology-Roorkee, India
Dhinaharan Nagamalai	Wireilla Net Solutions PTY LTD, Australia
Eric Renault	Institut Telecom - Telecom SudParis, Evry, France
Jacques Demerjian	Communication & Systems, France
James Henrydoss	AT&T and University of Colorado, USA
Krzysztof Walkowiak	Wroclaw University of Technology,Poland
Murugan D.	Manonmaniam Sundaranar University, India
Nabendu Chaki	University of Calcutta, India
Natarajan Meghanathan	Jackson State University, USA
Raja Kumar M.	Taylor's University, Malaysia
Salah Al-Majeed	University of Essex, UK
Selma Boumerdassi	Conservatoire National des Arts Et Metiers (CNAM), France
Sundarapandian Vaidyanathan	VelTech Dr. RR & Dr. SR Technical University, India

Program Committee Members

A.H.T. Mohammad	University of Bradford, UK
A.P. Sathish Kumar	PSG Institute of Advanced Studies, India
AAA. Atayero	Covenant University, Nigeria
Abdul Aziz	University of Central Punjab, Pakistan
Abdul Kadhir Ozcan	Karatay University, Turkey
Abdul Kadir Ozcan	The American University, Cyprus
Abdulbaset Mohammad	University of Bradford, United Kingdom
Ahmad Saad Al-Mogren	King Saud University, Saudi Arabia
Ahmed M. Khedr	Sharjah University, Sharjah, UAE
Ahmed Nada	Al-Quds University, Palestinian
Ajay K. Sharma	Dr. B R Ambedkar National Institute of Technology, India
Alaa Ismail Elnashar	Taif University, KSA
Alejandro Garces	Jaume I University, Spain
Alejandro Regalado Mendez	Universidad del Mar - México, USA
Alfio Lombardo	University of Catania, Italy
Ali El-Rashedy	University of Bridgeport, CT, USA
Ali M.	University of Bradford, United Kingdom
Ali Maqousi	Petra University, Jordan
Alireza Mahini	Islamic Azad University-Gorgan, Iran
Alvin Lim	Auburn University, USA
Amandeep Singh Thethi	Guru Nanak Dev University Amritsar, India
Amit Choudhary	Maharaja Surajmal Institute,India
Anand Sharma	MITS-Rajasthan, India
Anjan K.	RVCE-Bangalore, India
Ankit Thakkar	Nirma University, India
Ankit	BITS, PILANI India
Anthony Atayero	Covenant University, Nigeria
Aravind P.A.	Amrita School of Engineering India
Arun Pujari	Sambalpur University, India
Arunita Jaekel	University of Windsor, Canada
Ashok Kumar Das	IIT Hyderabad, India
Ashok kumar Sharma	YMCA Institute of Engineering, India
Ashutosh Dubey	NRI Institute of Science & Technology, Bhopal
Ashutosh Gupta	MJP Rohilkhand University, Bareilly
Athanasios Vasilakos	University of Western Macedonia, Greece
Azween Bin Abdullah	Universiti Teknologi Petronas, Malaysia
B. Srinivasan	Monash University, Australia
Babak Khosravifar	Concordia University, Canada
Balakannan S.P.	Chonbuk Nat. Univ., Jeonju
Balasubramanian K.	Lefke European University, Cyprus
Balasubramanian Karuppiah	Dr. MGR University, India
Bari A.	University of Western Ontario, Canada

Beatrice Cynthia Dhinakaran	TCIS, South Korea
Bela Genge	European Commission Joint Research Centre, Belgium
Bharat Bhushan Agarwal	I.F.T.M University, India
Bhupendra Suman	IIT Roorkee , India
Biju Pattnaik	University of Technology, India
Bikash singh	Islamic University-Kushtia, Bangladesh
Binod Kumar Pattanayak	Siksha O Anusandhan University, India
Bobby Barua	Ahsanullah University of Science and Technology, Bangladesh
Bong-Han	Kim, Chongju University, South Korea
Boo-Hyung Lee	KongJu National University, South Korea
Brajesh Kumar Kaushik	Indian Institute of Technology, India
Buket Barkana	University of Bridgeport, USA
Carlos E. Otero	University of South Florida Polytechnic, USA
Charalampos Z. Patrikakis	National Technical University of Athens, Greece
Chin-Chih Chang	Chung Hua University ,Taiwan
Cho Han Jin	Far East University, South Korea
Choudhari	Bhagwati Chaturvedi College of Engineering, India
Christos Politis	Kingston University, UK
Cristina Ribeiro	University of Waterloo, Canada
Cristina Serban	Ovidius University of Constantza, Romania
Danda B. Rawat	Old Dominion University, USA
David C. Wyld	Southeastern Louisiana University, USA
Debasis Giri	Haldia Institute of Technology, India
Debdatta Kandar	Sikkim Manipal University, India
Dhinaharan Nagamalai	Wirella Net Solutions PTY Ltd, Australia
Diego Reforgiato	University of Catania, Italy
Dimitris Kotzinos	Technical Educational Institution of Serres, Greece
Doreswamyh hosahalli	Mangalore University, India
Durga Toshniwal	Indian Institute of Technology, India
E. Martin	University of California, Berkeley, USA
E.V. Krishnamurthy	ANU College of Engg & Computer Science, Austraila
Emmanuel Bouix	iKlax Media, France
Eric Renault	Institut Telecom - Telecom SudParis, Evry, France
Ermatita Zuhairi	Sriwijaya University, Indonesia
Farag M. Sallabi	United Arab Emirates University, UAE
Farshad Safaei	Shahid Beheshti University, Iran
Ford Lumban Gaol	University of Indonesia
Genge Bela	Joint Research Centre, European Commission, Italy
Ghalem Belalem	University of Oran, Algeria
Giovanni Cordeiro Barroso	Universidade Federal do Ceara, Brasil
Giovanni Schembra	University of Catania, Italy
Girija Chetty	University of Canberra, Australia

Gomathi Kandasamy	Avinashilingam Deemed University for Women, India
Gopalakrishnan Kaliaperumal	Anna University, Chennai
Govardhan A.	JNTUH College of Engineering, India
Guo Bin	Institute TELECOM SudParis, France
H.V. Ramakrishnan	Dr. MGR University, India
Haider M. Alsabbagh	Basra University, Iraq
Haller Piroska	Petru Maior University-Tirgu Mures, Romania
Hao Shi	Victoria University, Australia
Hao-En Chueh	yuanpei University, Taiwan
Hari Chavan	National Institute of Technology, Jamshedpur, India
Henrique J.A. Holanda	UERN - Universidade do Estado do Rio Grande do Norte, Brasil
Henrique Joao Lopes Domingos	University of Lisbon, Portugal
Hiroyuki Hisamatsu	Osaka Electro-Communication University, Japan
Ho Dac Tu	Waseda University, Japan
Homam Reda El-Taj	Universiti Sains Malaysia, Malaysia
Hong yu	Capitol College, USA
Huosheng Hu	University of Essex, UK
Hussein Al-Bahadili	Petra University, Jordan
Hussein Ismail Khalaf Al-Bahadili	Petra University, Jordan
Hwangjun Song	Pohang University of Science and Technology,South Korea
Ignacio Gonzalez Alonso	University of Oviedo, Europe
Indrajit Bhattacharya	Kalyani Govt. Engg. College, India
Intisar Al-Mejibli	University of Essex, UK
Ioannis Karamitsos	Itokk Communications, Canada
J.K. Mandal	University of Kalyani, India
Jacques Demerjian	Communications & Systems, France
Jae Kwang Lee	Hannam University, South Korea
Jalel Akaichi	University of Tunis, Tunisia
Jan Zizka	SoNet/DI, FBE, Mendel University in Brno, Czech Republic
Jeong-Hyun Park	Electronics Telecommunication Research Institute, South Korea
Jeyanthy N.	VIT University, India
Jifeng Wang	University of Illinois at Urbana Champaign, USA
Johann Groschdl	University of Bristol, UK
Jose Enrique Armendariz-Inigo	Universidad Publica de Navarra, Spain
Juan Li	North Dakota State University, USA
Jyoti Singhai	Electronics and Communication Deptt-MANIT, India
Jyotirmay Gadewadikar	Alcorn State University, USA
Kai Xu	University of Bradford, United Kingdom
Kamalrulnizam Abu Bakar	Universiti Teknologi Malaysia, Malaysia

Karim Konate	University Cheikh Anta DIOP, Dakar
Kaushik Chakraborty	Jadavpur University, India
Kayhan Erciyes	Izmir University, Turkey
Khaled Shuaib	United Arab Emirates University, UAE
Khamish Malhotra	University of Glamorgan, UK
Khoa N. Le	University of Western Sydney, Australia
Krishnamurthy E.V.	ANU College of Engg & Computer Science, Austraila
Krzysztof Walkowiak	Wroclaw University of Technology, Poland
Kuribayashi	Seikei University, Japan
L. Nirmala Devi	Osmania University - Hyderabad, India
Laiali Almazaydeh	University of Bridgeport, USA
Lu Yan	University of Hertfordshire, UK
Lus Veiga	Technical University of Lisbon, Portugal
Lylia Abrouk	University of Burgundy, France
M. Aqeel Iqbal	FUIEMS, Pakistan
M. Rajarajan	City University, UK
M. Ali	University of Bradford, UK
Maode Ma	Nanyang Technological University, Singapore
Marco Folli	University of Pavia, Italy
Marco Roccetti	Universty of Bologna, Italy
Massimo Esposito	ICAR-CNR, Italy
Md. Sipon Miah	Islamic University-Kushtia, Bangladesh
Michal Wozniak	Wroclaw University of Technology, Poland
Michel Owayjan	American University of Science & Technology, Lebanon
Miguel A. Wister	Juarez Autonomous University of Tabasco, Mexico
Mohamed Hassan	American University of Sharjah, UAE
Mohammad Ali Jabreil Jamali	Islamic Azad University, Iran
Mohammad Hadi Zahedi	Ferdowsi University of Mashhad, Iran
Mohammad Hajjar	Lebanese University, Lebanon
Mohammad Kaghazgaran	Islamic Azad University, Iran
Mohammad Mehdi Farhangia	Universiti Teknologi Malaysia, Malaysian
Mohammad Momani	University of technology Sydney, Australia
Mohammad Talib	University of Botswana, Botswana
Mohammad Zaidul Karim	Daffodil International University, Bangladesh
Mohammed Feham	University of Tlemcen, Algeria
Mohammed M. Alkhawlani	University of Science and Technology, Yemen
Mohsen Sharifi	Iran University of Science and Technology, Iran
Muhammad Sajjadur Rahim	University of Rajshahi, Bangladesh
Murty	Ch A S, JNTU, Hyderabad
Murugan D.	Manonmaniam Sundaranar University, India
Mydhili Nair	M S Ramaiah Institute of Technology, India
N. Krishnan	Manonmaniam Sundaranar University, India
Nabendu Chaki	University of Calcutta, India

Nadine Akkari	King abdulaziz University, Saudi Arabia
Naohiro Ishii	Aichi Institute of Technology, Japan
Nasrollah M. Charkari	Tarbiat Modares University, Iran
Natarajan Meghanathan	Jackson State University, USA
Nicolas Sklavos	Technological Educational Institute of Patras, Greece
Nidaa Abdual Muhsin Abbas	University of Babylon, Iraq
Nour Eldin Elmadany	Arab Acadmy for Science and Technology, Egypt
Ognjen Kuljaca	Alcorn State University, USA
Olakanmi Oladayo	University of Ibadan, Nigeria
Omar Almomani	Universiti Utara Malaysia, Malaysia
Orhan Dagdeviren	Izmir University, Turkey
Osman B. Ghazali	Universiti Utara Malaysia, Malaysia
Othon Marcelo Nunes Batista	Universidade Salvador, Brazil
Padmalochan Bera	Indian Institute of Technology, Kharagpur, India
Partha Pratim Bhattacharya	Mody Institute of Technology & Science, India
Patricia Marcu	Leibniz Supercomputing Centre, Germany
Patrick Seeling	University of Wisconsin - Stevens Point, USA
R. Thandeeswaran	VIT University, India
Phan Cong Vinh	London South Bank University, UK
Pinaki Sarkar	Jadavpur University, India
Polgar Zsolt Alfred	Technical University of Cluj Napoca, Romania
Ponpit Wongthongtham	Curtin University of Technology, Australia
Quan (Alex) Yuan	University of Wisconsin-Stevens Point, USA
Rafael Timoteo	University of Brasilia - UnB, Brazil
Raied Salman	Virginia Commonwealth University, USA
Rajendra Akerkar	Technomathematics Research Foundation, India
Rajeswari Balasubramaniam	Dr. MGR University, India
Rajkumar Kannan	Bishop Heber College, India
Rakhesh Singh Kshetrimayum	Indian Institute of Technology, Guwahati, India
Raman Maini	Punjabi University, India
Ramayah Thurasamy	Universiti Sains Malaysia, Malaysia
Ramayah	Universiti Sains Malaysia, Malaysia
Ramin karimi	University Technology Malaysia
Razvan Deaconescu	University Politehnica of Bucharest, Romania
Reena Dadhich	Govt. Engineering College Ajmer
Reshmi Maulik	University of Calcutta, India
Reza Ebrahimi Atani	University of Guilan, Iran
Rituparna Chaki	West Bengal University of Technology, India
Robert C. Hsu	Chung Hua University, Taiwan
Roberts Masillamani	Hindustan University, India
Rohitha Goonatilake	Texas A&M International University, USA
Rushed Kanawati	LIPN - Universite Paris 13, France
S. Geetha	Anna University - Tiruchirappalli, India
S. Hariharan	B.S. Abdur Rahman University, India

S. Venkatesan	University of Texas at Dallas - Richardson, USA
S.A.V. Satyamurty	Indira Gandhi Centre for Atomic Research, India
S. Arivazhagan	Mepco Schlenk Engineering College, India
S. Li	Swansea University, UK
S. Senthil Kumar	Universiti Sains Malaysia, Malaysia
Sajid Hussain	Acadia University, Canada
Salah M. Saleh Al-Majeed	University of Essex, United Kingdom
Saleena Ameen	B.S.Abdur Rahman University, India
Salem Nasri	ENIM, Monastir University, Tunisia
Salim Lahmiri	University of Qubec at Montreal, Canada
Salini P.	Pondichery Engineering College, India
Salman Abdul Moiz	Centre for Development of Advanced Computing, India
Samarendra Nath Sur	Sikkim Manipal University, India
Sami Ouali	ENSI, Compus of Manouba, Manouba, Tunisia
Samiran Chattopadhyay	Jadavpur University, India
Samodar reddy	India school of mines , India
Samuel Falaki	Federal University of Technology-Akure, Nigeria
Sanjay Singh	Manipal Institute of Technology, India
Sara Najafzadeh	University Technology Malaysia
Sarada Prasad Dakua	IIT-Bombay, India
Sarmistha Neogy	Jadavpur University, India
Satish Mittal	Punjabi University, India
S.C. SHARMA	IIT - Roorkee, India
Seetha Maddala	CBIT, Hyderabad
Selma Boumerdassi	Cnam/Cedric, France
Sergio Ilarri	University of Zaragoza, Spain
Serguei A. Mokhov	Concordia University, Canada
Shaoen Wu	The University of Southern Mississippi, USA
Sharvani G.S.	RV College of Engineering, Inida
Sherif S. Rashad	Morehead State University, USA
Shin-ichi Kuribayashi	Seikei University, Japan
Shivan Haran	Arizona state University, USA
Shobha Shankar	Vidya vardhaka College of Engineering, India
Shrikant K. Bodhe	Bosh Technologies, India
Shriram Vasudevan	VIT University, India
Shrirang Ambaji Kulkarni	National Institute of Engineering, India
Shubhamoy Dey	Indian Institute of Management Indore, India
Solange Rito Lima	University of Minho, Portugal
Souad Zid	National Engineering School of Tunis, Tunisia
Soumyabrata Saha	Guru Tegh Bahadur Institute of Technology, India
Sridharan	CEG Campus - Anna University, India
Sriman Narayana Iyengar	VIT University, India
Srinivasulu Pamidi	V R Siddhartha Engineering College Vijayawada, India

Sriram Maturi	Osmania University, India
Subhabrata Mukherjee	Jadavpur University, India
Subir Sarkar	Jadavpur University, India
Sundarapandian Vaidyanathan	VelTech Dr. RR & Dr. SR Technical University, India
Sunil Singh	Bharati vidyapeeth's College of Engineering, India
Sunilkumar S. Manvi	REVA Institute of Technology and Management Kattigenhalli, India
SunYoung Han	Konkuk University, South Korea
Susana Sargento	University of Aveiro, Portugal
Swarup Mitra	Jadavpur University, Kolkata, India
T. Ambaji Venkat Narayana Rao	Hyderabad Institution of Technology and Management , India
T.G. Basavaraju	National Institute of Technology Karnataka (NITK), India
Thomas Yang	Embry Riddle Aeronautical University, USA
Tri Kurniawan Wijaya	Technische Universitat Dresden, Germany
Tsung Teng Chen	National Taipei Univ., Taiwan
Utpal Biswas	University of Kalyani, India
V.M. Pandharipande	Dr. Babasaheb Ambedkar Marathwada University, India
Valli Kumari Vatsavayi	AU College of Engineering, India
Vijayalakshmi S.	VIT University, India
Virgil Dobrota	Technical University of Cluj-Napoca, Romania
Vishal Sharma	Metanoia Inc., USA
Wei Jie	University of Manchester, UK
Wichian Sittiprapaporn	Mahasarakham University, Thailand
Wided Oueslati	l'institut Superieur de Gestion de Tunis, Tunisia
William R. Simpson	Institute for Defense Analyses, USA
Wojciech Mazurczyk	Warsaw University of Technology, Poland
Xiaohong Yuan	North Carolina A & T State University, USA
Xin Bai	The City University of New York, USA
Yahya Slimani	Faculty of Sciences of Tunis, Tunisia
Yannick Le Moullec	Aalborg University, Denmark
Yaser M. Khamayseh	Jordan University of Science and Technology, Jordan
Yedehalli Kumara Swamy	Dayanand Sagar College of Engineering, India
Yeong Deok Kim	Woosong University, South Korea
Yogeshwar Kosta	Marwadi Education Foundations Group of Institutions, India
Yuh-Shyan Chen	National Taipei University, Taiwan
Yung-Fa Huang	Chaoyang University of Technology, Taiwan
Zaier Aida	National Engeneering School of GABES, Tunisia
Zakaria Moudam	Université sidi mohammed ben Abdellah, Morocco
Zuqing Zhu	Cisco Systems, USA

External Reviewers

A. Kannan	K.L.N. College of Engineering, India
Martin	Sri Manakula Vinayagar Engineering College, India
Abhishek Samanta	Jadavpur University, Kolkata, India
Ayman Khalil	Institute of Electronics and Telecommunications of Rennes, France
Cauvery Giri	RVCE, India
Ch. V. Rama Rao	Gudlavalleru Engineering College, India
Chandra Mohan	Bapatla Engineering College, India
E.P. Ephzibah	VIT University-Vellore, India
Hameem Shanavas	Vivekananda Institute of Technolgy, India
Kota Sunitha	G. Narayanamma Institute of Technology and Science, Hyderabad
Kunjal B. Mankad	ISTAR, Gujarat, India
Lakshmi Rajamani	Osmania University, India
Lavanya	Blekinge Institute of Technology, Sweden
M.P. Singh	National Institute of Technology, Patna
M. Tariq Banday	University of Kashmir, India
M.M.A. Hashem	Khulna University of Engineering and Technology, Bangladesh
Mahalinga V. Mandi	Dr. Ambedkar Institute of Technology, Bangalore, Karnataka, India
Mahesh Goyani	G H Patel College of Engineering and Technology, India
Maragathavalli P.	Pondicherry Engineering College, India
M.P. Singh	National Institute of Technology, Patna
M. Tariq Banday	University of Kashmir, India
M.M.A. Hashem	Khulna University of Engineering and Technology, Bangladesh
Mahalinga V. Mandi	Dr. Ambedkar Institute of Technology, India
Monika Verma	Punjab Technical University, India
Moses Ekpenyong	University of Uyo, Nigeria
Mini Patel	Malwa Institute of Technology, India
N. Kaliammal	NPR College of Engg &Tech, India
N. Adhikari	Biju Pattnaik University of Technology, India
N.K. Choudhari	Bhagwati Chaturvedi College of Engineering, India
Naga Prasad Bandaru	PVP Siddartha Institute of Technology, India
Nagamanjula Prasad	Padmasri Institute of Technology, India
Nagaraj Aitha	I.T, Kamala Institute of Tech & Science, India
Nana Patil	NIT Surat, Gujrat
Nitiket N. Mhala	B.D. College of Engineering - Sewagram, India
P. Ashok Babu	Narsimhareddy Engineering College, India
P. Sheik Abdul Khader	B.S. Abdur Rahman University, India

Pesn Krishna Prasad	Aditya Engineering College, India
Partha Pratim Bhattacharya	Mody Institute of Technology & Science, India
Pappa Rajan	Anna University, India
Pooja Jain	JUIT Waknaghat, India
Prabu Dorairaj	NetApp Inc, India
Pradeepini Gera	Jawaharlal Nehru Technological University, India
Pravin P. Karde	HVPM's College of Engineering & Technology - Amravati, India
Premanand K. Kadbe	Vidya Pratishthan's College of Engineering, India
Priti Sajja	Sardar Patel University, India
R. Baskaran	Anna University - Chennai, India
R. Murali	Dr. Ambedkar Institute of Technology, Bangalore
R.M. Suresh	Mysore University
Rabindranath Bera	Sikkim Manipal Inst. of Technol., India
Rajashree Biradar	Ballari Institute of Technology and Management, India
Rajesh Kumar Krishnan	Bannari Amman Inst. of Technol., India
Rajesh Kumar P.	The Best International, Australia
Rajeshwari Hegde	BMS College of Engineering, India
Rajagopal Palsonkennedy	Dr. MGR University, India
Selvakumar Ramachandran	Blekinge Institute of Technology, Sweden

Contents

Computing and Information Technology Algorithms and Applications

Digital Image Processing and Pattern Recognition

Analysis, Control and Synchronization of Hyperchaotic Zhou System via Adaptive Control

Sundarapandian Vaidyanathan

R & D Centre, Vel Tech Dr. RR & Dr. SR Technical University
Avadi-Alamathi Road, Avadi, Chennai-600 062, India
sundarvtu@gmail.com
http://www.vel-tech.org/

Abstract. This paper investigates the analysis, control and synchronization of the hyperchaotic Zhou system (2009) via adaptive control. First, an adaptive control scheme is derived to stabilize the hyperchaotic Zhou system with unknown parameters to its unstable equilibrium at the origin. Then an adaptive synchronization scheme is derived to achieve global chaos synchronization of the identical hyperchaotic Zhou systems with unknown parameters. The results derived for adaptive stabilization and synchronization for the hyperchaotic system are established using the Lyapunov stability theory. Numerical simulations are shown to demonstrate the effectiveness of the adaptive control and synchronization schemes derived in this paper.

Keywords: Adaptive control, hyperchaos, synchronization, hyperchaotic Zhou system.

1 Introduction

Hyperchaotic system is defined as a chaotic system with more than one positive Lyapunov exponent. Hyperchaotic system has the characteristics of high capacity, high security and high efficiency. Typical examples of hyperchaotic systems are hyperchaotic Rössler system [1], hyperchaotic Lorenz-Haken system [2] and hyperchaotic Chua's circuit [3].

The problem of controlling a chaotic system was introduced by Ott *et al.* ([4], 1990). The control of chaotic systems is basically to design state feedback control laws that stabilizes the chaotic systems around the unstable equilibrium points. Active control method is used when the system parameters are known and adaptive control method is used when some or all of the system parameters are unknown ([4]-[6]).

Chaos synchronization is a phenomenon that may occur when two or more chaotic oscillators are coupled or when a chaotic oscillator drives another chaotic oscillator. In most of the chaos synchronization approaches, the *master-slave* or *drive-response* formalism is used. If a particular chaotic system is called the *master* or *drive* system and another chaotic system is called the *slave* or *response* system, then the idea of chaos synchronization is to use the output of the master system to control the slave system so that the output of the slave system tracks the output of the master system asymptotically.

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 1–10.
springerlink.com

Since the pioneering work by Pecora and Carroll ([7], 1990), several approaches have been proposed for chaos synchronization such as the active control method ([8]-[9]), adaptive control method ([10]-[12]), sampled-data control method [13], backstepping method [14], sliding mode control method ([15]-[16]), etc.

This paper investigates the analysis, control and synchronization for the hyperchaotic Zhou system (Zhou *et al.* [17], 2009). First, we derive adaptive feedback control for the hyperchaotic Zhou system about its unstable equilibrium at the origin. Then we derive adaptive synchronization scheme for the identical hyperchaotic Zhou systems. The adaptive control and synchronization results derived in this paper are established using Lyapunov stability theory [18].

This paper has been organized as follows. In Section 2, we give a system description and qualitative analysis of the hyperchaotic Zhou system. In Section 3, we derive results for the adaptive control of the hyperchaotic Zhou system with unknown parameters. In Section 4, we derive results for the adaptive synchronization of the identical hyperchaotic Zhou systems with unknown parameters. In Section 5, we summarize the main results obtained in this paper.

2 Analysis of the Hyperchaotic Zhou System

The hyperchaotic Zhou system ([17], 2009) is described by the 4D dynamics

$$\begin{aligned} \dot{x}_1 &= a(x_2 - x_1) + x_4 \\ \dot{x}_2 &= cx_2 - x_1x_3 \\ \dot{x}_3 &= -bx_3 + x_1x_2 \\ \dot{x}_4 &= dx_1 + 0.5x_2x_3 \end{aligned} \tag{1}$$

where x_1, x_2, x_3, x_4 are the state variables of the system and a, b, c, d are constant, positive parameters of the system.

The system (1) is symmetrical about the x_3-axis because it is invariant under the coordinate transformation

$$(x_1, x_2, x_3, x_4) \to (-x_1, -x_2, x_3, -x_4)$$

and the phase portrait of the system (1) in $x_1x_2x_4$-three dimensional space is symmetrical about the origin.

The system (1) is *hyperchaotic* when

$$a = 35, \quad b = 3, \quad c = 12 \ \text{ and } \ 0 < d \leq 34.8 \tag{2}$$

Figure 1 describes the phase portrait of the hyperchaotic system (1) where the parameters are takes as in (2) with $d = 1$.

Obviously, the hyperchaotic system (1) has only an equilibrium point at the origin.

The linearization matrix of the system (1) at the origin is given by

$$A = \begin{bmatrix} -a & a & 0 & 1 \\ 0 & c & 0 & 0 \\ 0 & 0 & -b & 0 \\ d & 0 & 0 & 0 \end{bmatrix}$$

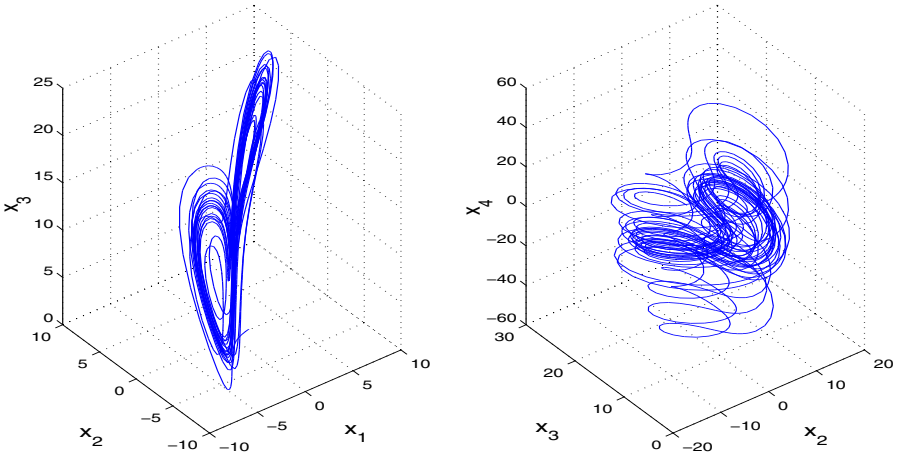

Fig. 1. Phase Portrait of the Hyperchaotic Zhou System

The eigenvalues of A are

$$\lambda_1 = c, \lambda_2 = -b, \lambda_3 = \frac{-a + \sqrt{a^2 + 4d}}{2}, \lambda_4 = \frac{-a - \sqrt{a^2 + 4d}}{2}$$

Since $\lambda_1 = c > 0$, it follows by the Lyapunov stability theory [18] that the origin is an unstable equilibrium of the system (1).

3 Adaptive Control of the Hyperchaotic Zhou System

3.1 Main Results

In this section, we discuss the adaptive controller design for globally stabilizing the hyperchaotic Zhou system (2009), when the parameter values are unknown.

Thus, we consider the controlled hyperchaotic Zhou system described by the dynamics

$$\begin{aligned} \dot{x}_1 &= a(x_2 - x_1) + x_4 + u_1 \\ \dot{x}_2 &= cx_2 - x_1x_3 + u_2 \\ \dot{x}_3 &= -bx_3 + x_1x_2 + u_3 \\ \dot{x}_4 &= dx_1 + 0.5x_2x_3 + u_4 \end{aligned} \tag{3}$$

where u_1, u_2, u_3, u_4 are feedback controllers to be designed using the states x_1, x_2, x_3, x_4 and estimates $\hat{a}, \hat{b}, \hat{c}, \hat{d}$ of the unknown system parameters a, b, c, d of the system.

Next, we consider the following adaptive control functions

$$\begin{aligned} u_1 &= -\hat{a}(x_2 - x_1) - x_4 - k_1x_1 \\ u_2 &= -\hat{c}x_2 + x_1x_3 - k_2x_2 \\ u_3 &= \hat{b}x_3 - x_1x_2 - k_3x_3 \\ u_4 &= -\hat{d}x_1 - 0.5x_2x_3 - k_4x_4 \end{aligned} \tag{4}$$

where $\hat{a}, \hat{b}, \hat{c}$ and $\hat{d}$ are the estimates of the system parameters a, b, c and d, respectively, and $k_i, (i = 1, 2, 3, 4)$ are positive constants.

Substituting the control law (4) into the plant equation (3), we obtain

$$\begin{aligned} \dot{x}_1 &= (a - \hat{a})(x_2 - x_1) - k_1 x_1 \\ \dot{x}_2 &= (c - \hat{c}) x_2 - k_2 x_2 \\ \dot{x}_3 &= -(b - \hat{b}) x_3 - k_3 x_3 \\ \dot{x}_4 &= (d - \hat{d}) x_1 - k_4 x_4 \end{aligned} \tag{5}$$

We define the parameter estimation error as

$$e_a = a - \hat{a}, \quad e_b = b - \hat{b}, \quad e_c = c - \hat{c}, \quad e_d = d - \hat{d} \tag{6}$$

Using (6), the state dynamics (3) can be simplified as

$$\begin{aligned} \dot{x}_1 &= e_a(x_2 - x_1) - k_1 x_1 \\ \dot{x}_2 &= e_c x_2 - k_2 x_2 \\ \dot{x}_3 &= -e_b x_3 - k_3 x_3 \\ \dot{x}_4 &= e_d x_1 - k_4 x_4 \end{aligned} \tag{7}$$

We use Lyapunov approach for the derivation of the update law for adjusting the parameter estimates $\hat{a}, \hat{b}, \hat{c}$ and $\hat{d}$.

Consider the quadratic Lyapunov function defined by

$$V(x_1, x_2, x_3, x_4, e_a, e_b, e_c, e_d) = \frac{1}{2}\left(x_1^2 + x_2^2 + x_3^2 + x_4^2 + e_a^2 + e_b^2 + e_c^2 + e_d^2\right) \tag{8}$$

which is a positive definite function on $\mathbb{R}^8$.

Note that

$$\dot{e}_a = -\dot{\hat{a}}, \quad \dot{e}_b = -\dot{\hat{b}}, \quad \dot{e}_c = -\dot{\hat{c}}, \quad \dot{e}_d = -\dot{\hat{d}} \tag{9}$$

Differentiating V along the trajectories of (5) and using (9), we obtain

$$\begin{aligned} \dot{V} = -k_1 x_1^2 - k_2 x_2^2 - k_3 x_3^2 - k_4 x_4^2 + e_a\left[x_1(x_2 - x_1) - \dot{\hat{a}}\right] \\ + e_b\left[-x_3^2 - \dot{\hat{b}}\right] + e_c\left[x_2^2 - \dot{\hat{c}}\right] + e_d\left[x_1 x_4 - \dot{\hat{d}}\right] \end{aligned} \tag{10}$$

In view of Eq. (10), the estimated parameters are updated by the following law:

$$\begin{aligned} \dot{\hat{a}} &= x_1(x_2 - x_1) + k_5 e_a \\ \dot{\hat{b}} &= -x_3^2 + k_6 e_b \\ \dot{\hat{c}} &= x_2^2 + k_7 e_c \\ \dot{\hat{d}} &= x_1 x_4 + k_8 e_d \end{aligned} \tag{11}$$

where k_5, k_6, k_7 and k_8 are positive constants.

Next, we prove the following result.

Theorem 1. *The hyperchaotic Zhou system (3) with unknown parameters is globally and exponentially stabilized by the adaptive control law (4), where the update law for the parameters is given by (11) and $k_i, (i = 1, 2, \ldots, 8)$ are positive constants.*

Proof. Substituting (11) into (10), we obtain

$$\dot{V} = -k_1 x_1^2 - k_2 x_2^2 - k_3 x_3^2 - k_4 x_4^2 - k_5 e_a^2 - k_6 e_b^2 - k_7 e_c^2 - k_8 e_d^2 \qquad (12)$$

which is a negative definite function on $\mathbb{R}^8$.

Thus, by Lyapunov stability theory [18], it follows that the plant dynamics (7) is globally exponentially stable and also that the parameter estimate errors e_a, e_b, e_c, e_d converge to zero exponentially with time. □

3.2 Numerical Results

For the simulations, the fourth order Runge-Kutta method with step-size $h = 10^{-8}$ is used to solve the hyperchaotic Zhou system (3) with the adaptive control law (4) and the parameter update law (11). The parameters of the system (3) are selected as $a = 35,\ b = 3,\ c = 12$ and $d = 1$. We also take $k_i = 4$ for $i = 1, 2, \ldots, 8$.

Suppose that the initial values of the estimated parameters are

$$\hat{a}(0) = 8,\ \ \hat{b}(0) = 24,\ \ \hat{c}(0) = 30,\ \ \ \hat{d}(0) = 17$$

Suppose that we take the initial values of the states of the system (3) as

$$x_1(0) = -6,\ \ x_2(0) = 7,\ \ x_3(0) = 20,\ \ \ x_4(0) = -18$$

Figure 2 shows that the states of the closed-loop system (7) converge to the equilibrium $E_0 = (0, 0, 0, 0)$ exponentially with time. Figure 3 shows that the estimates $\hat{a}, \hat{b}, \hat{c}, \hat{d}$ converge to the system parameters a, b, c, d exponentially with time.

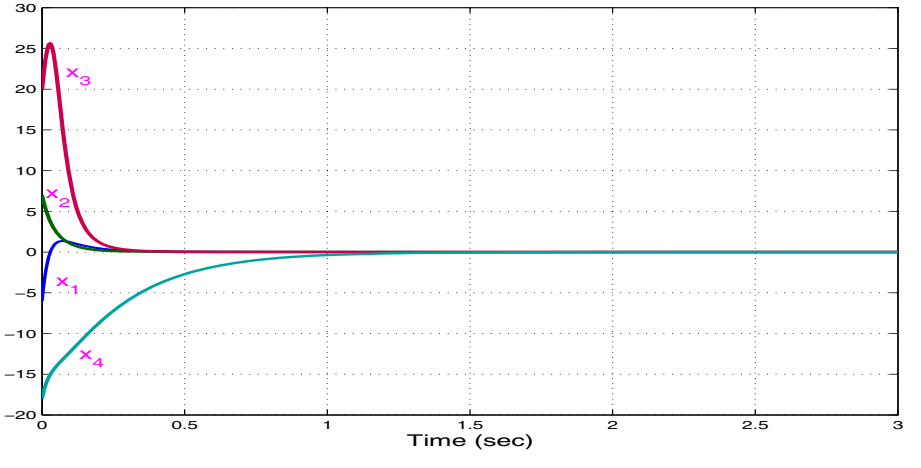

Fig. 2. Time History of the States of the Controlled Hyperchaotic Zhou System

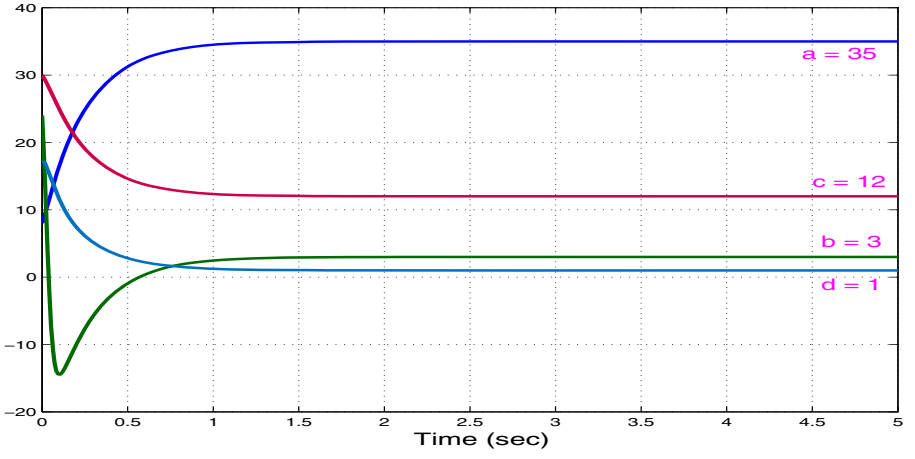

Fig. 3. Time History of the Parameter Estimates $\hat{a}(t), \hat{b}(t), \hat{c}(t), \hat{d}(t)$

4 Adaptive Synchronization of the Hyperchaotic Zhou System

4.1 Main Results

In this section, we discuss the adaptive synchronizer design for the identical hyperchaotic Zhou systems (2009) with unknown parameters.

As the master system, we consider the hyperchaotic Zhou dynamics described by

$$\begin{aligned} \dot{x}_1 &= a(x_2 - x_1) + x_4 \\ \dot{x}_2 &= cx_2 - x_1x_3 \\ \dot{x}_3 &= -bx_3 + x_1x_2 \\ \dot{x}_4 &= dx_1 + 0.5x_2x_3 \end{aligned} \tag{13}$$

where x_1, x_2, x_3, x_4 are the state variables and a, b, c, d are unknown system parameters.

As the slave system, we consider the controlled hyperchaotic Zhou dynamics described by

$$\begin{aligned} \dot{y}_1 &= a(y_2 - y_1) + y_4 + u_1 \\ \dot{y}_2 &= cy_2 - y_1y_3 + u_2 \\ \dot{y}_3 &= -by_3 + y_1y_2 + u_3 \\ \dot{y}_4 &= dy_1 + 0.5y_2y_3 + u_4 \end{aligned} \tag{14}$$

where y_1, y_2, y_3 are the state variables and u_1, u_2, u_3 are the nonlinear controllers to be designed.

The synchronization error e is defined by

$$e_i = y_i - x_i, \quad (i = 1, 2, 3, 4) \tag{15}$$

Then the error dynamics is obtained as

$$\begin{aligned}\dot{e}_1 &= a(e_2 - e_1) + e_4 + u_1\\ \dot{e}_2 &= ce_2 - y_1y_3 + x_1x_3 + u_2\\ \dot{e}_3 &= -be_3 + y_1y_2 - x_1x_2 + u_3\\ \dot{e}_4 &= de_1 + 0.5(y_2y_3 - x_2x_3) + u_4\end{aligned} \tag{16}$$

We define the adaptive synchronizing law

$$\begin{aligned}u_1 &= -\hat{a}(e_2 - e_1) - e_4 - k_1e_1\\ u_2 &= -\hat{c}e_2 + y_1y_3 - x_1x_3 - k_2e_2\\ u_3 &= \hat{b}e_3 - y_1y_2 + x_1x_2 - k_3e_3\\ u_4 &= -\hat{d}e_1 - 0.5(y_2y_3 - x_2x_3) - k_4e_4\end{aligned} \tag{17}$$

where $\hat{a}, \hat{b}, \hat{c}$ and $\hat{d}$ are estimates of the system parameters a, b, c and d, respectively, and $k_i, (i = 1, 2, 3, 4)$ are positive constants.

Substituting (17) into (16), we obtain the closed-loop error dynamics as

$$\begin{aligned}\dot{e}_1 &= (a - \hat{a})(e_2 - e_1) - k_1e_1\\ \dot{e}_2 &= (c - \hat{c})e_2 - k_2e_2\\ \dot{e}_3 &= -(b - \hat{b})e_3 - k_3e_3\\ \dot{e}_4 &= (d - \hat{d})e_1 - k_4e_4\end{aligned} \tag{18}$$

We define the parameter estimation error as

$$e_a = a - \hat{a}, \quad e_b = b - \hat{b}, \quad e_c = c - \hat{c} \text{and} \quad e_d = d - \hat{d} \tag{19}$$

Substituting (19) into (18), the error dynamics (18) can be simplified as

$$\begin{aligned}\dot{e}_1 &= e_a(e_2 - e_1) - k_1e_1\\ \dot{e}_2 &= e_ce_2 - k_2e_2\\ \dot{e}_3 &= -e_be_3 - k_3e_3\\ \dot{e}_4 &= e_de_1 - k_4e_4\end{aligned} \tag{20}$$

We use Lyapunov approach for the derivation of the update law for adjusting the parameter estimates $\hat{a}$, $\hat{b}$, $\hat{c}$ and $\hat{d}$.

Consider the quadratic Lyapunov function defined by

$$V(e_1, e_2, e_3, e_4, e_a, e_b, e_c, e_d) = \frac{1}{2}\left(e_1^2 + e_2^2 + e_3^2 + e_4^2 + e_a^2 + e_b^2 + e_c^2 + e_d^2\right) \tag{21}$$

which is a positive definite function on $\mathbb{R}^8$.

Note that

$$\dot{e}_a = -\dot{\hat{a}}, \quad \dot{e}_b = -\dot{\hat{b}}, \quad \dot{e}_c = -\dot{\hat{c}}, \quad \dot{e}_d = -\dot{\hat{d}}. \tag{22}$$

Differentiating V along the trajectories of (20) and using (22), we obtain

$$\begin{aligned}\dot{V} = -k_1e_1^2 - k_2e_2^2 - k_3e_3^2 - k_4e_4^2 + e_a\left[e_1(e_2 - e_1) - \dot{\hat{a}}\right]\\ +e_b\left[-e_3^2 - \dot{\hat{b}}\right] + e_c\left[e_2^2 - \dot{\hat{c}}\right] + e_d\left[e_1e_4 - \dot{\hat{d}}\right]\end{aligned} \tag{23}$$

In view of Eq. (23), the estimated parameters are updated by the following law:

$$\begin{aligned}\dot{\hat{a}} &= e_1(e_2 - e_1) + k_5 e_a \\ \dot{\hat{b}} &= -e_3^2 + k_6 e_b \\ \dot{\hat{c}} &= e_2^2 + k_7 e_c \\ \dot{\hat{d}} &= e_1 e_4 + k_8 e_d\end{aligned} \tag{24}$$

where k_5, k_6, k_7 and k_8 are positive constants.

Theorem 2. *The identical hyperchaotic Zhou systems (13) and (14) with unknown parameters are globally and exponentially synchronized by the adaptive control law (17), where the update law for the parameters is given by (24) and $k_i, (i = 1, 2, \ldots, 6)$ are positive constants.*

Proof. Substituting (24) into (23), we obtain

$$\dot{V} = -k_1 e_1^2 - k_2 e_2^2 - k_3 e_3^2 - k_4 e_4^2 - k_5 e_a^2 - k_6 e_b^2 - k_7 e_c^2 - k_8 e_d^2 \tag{25}$$

which is a negative definite function on $\mathbb{R}^8$.

Thus, by Lyapunov stability theory [18], it follows that the error dynamics (20) is globally exponentially stable and also that the parameter estimate errors e_a, e_b, e_c, e_d converge to zero exponentially with time. □

4.2 Numerical Results

For the numerical simulations, the fourth order Runge-Kutta method with step-size $h = 10^{-8}$ is used to solve the hyperchaotic Zhou systems (13) and (14) with the adaptive control law (17) and the parameter update law (24).

The parameters of the hyperchaotic Zhou systems are selected as $a = 35, b = 3, c = 12$ and $d = 1$.

We also take $k_i = 4$ for $i = 1, 2, \ldots, 8$.

Suppose that the initial values of the estimated parameters are

$$\hat{a}(0) = 7, \;\; \hat{b}(0) = 16, \;\; \hat{c}(0) = 20, \;\; \hat{d}(0) = 9$$

Suppose that the initial values of the master system (13) are taken as

$$x_1(0) = 7, \;\; x_2(0) = -14, \;\; x_3(0) = 12, \;\; x_4(0) = -17$$

Suppose that the initial values of the slave system (14) are taken as

$$y_1(0) = -8, \;\; y_2(0) = 25, \;\; y_3(0) = 9, \;\; y_4(0) = 12$$

Figure 4 shows that the identical hyperchaotic Zhou systems (13) and (14) are exponentially synchronized with time. Figure 5 shows that the parameter estimates $\hat{a}, \hat{b}, \hat{c}, \hat{d}$ converge to the system parameters a, b, c, d exponentially with time.

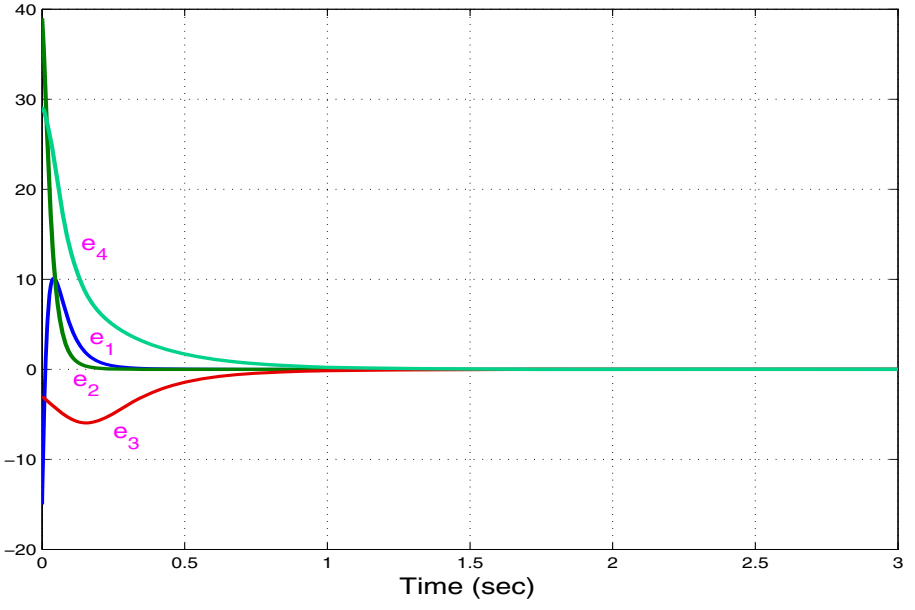

Fig. 4. Time History of the Synchronization Error for Identical Hyperchaotic Zhou Systems

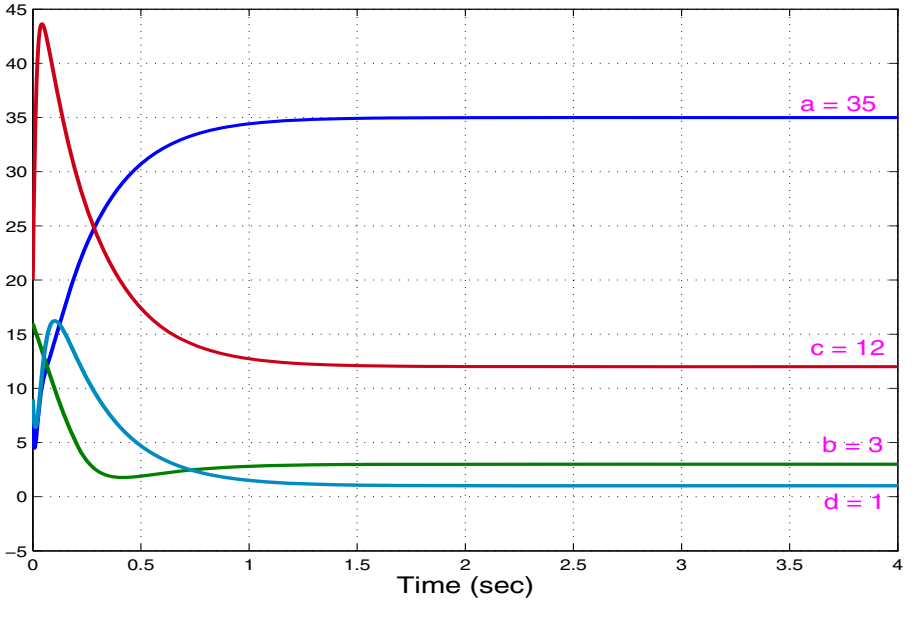

Fig. 5. Time History of the Parameter Estimates $\hat{a}(t), \hat{b}(t), \hat{c}(t), \hat{d}(t)$

5 Conclusions

In this paper, we derived results for the adaptive controller and synchronizer for the hyperchaotic Zhou system (Zhou *et al.* 2009) with unknown parameters. First, we designed an adaptive control scheme to stabilize the hyperchaotic Zhou system to its unstable equilibrium point at the origin based on the Lyapunov stability theory. Then we designed an adaptive synchronization scheme for the global chaos synchronization of the identical hyperchaotic Zhou systems with unknown parameters. Our synchronization results were established using the Lyapunov stability theory. Numerical simulations are presented to demonstrate the effectiveness of the adaptive controller and synchronizer schemes derived for the hyperchaotic Zhou system (2009).

References

1. Rössler, O.E.: An equation for hyperchaos. Phys. Lett. A 71, 155–157 (1979)
2. Ning, C.Z., Haken, H.: Detuned lasers and the complex Lorenz equations: Subcritical and supercritical Hopf bifurcations. Phys. Rev. A 41, 3826–3837 (1990)
3. Kapitaniak, T., Chua, L.O.: Hyperchaotic attractor of unidirectionally coupled Chua's circuit. Int. J. Bifurcat. Chaos 4, 477–482 (1994)
4. Ott, E., Grebogi, C., Yorke, J.A.: Controlling chaos. Phys. Rev. Lett. 64, 1196–1199 (1990)
5. Ge, S.S., Wang, C., Lee, T.H.: Adaptive backstepping control of a class of chaotic systems. Internat. J. Bifur. Chaos 10, 1149–1156 (2000)
6. Sun, M., Tian, L., Jiang, S., Xun, J.: Feedback control and adaptive control of the energy resource chaotic system. Chaos, Solitons & Fractals 32, 168–180 (2007)
7. Pecora, L.M., Carroll, T.L.: Synchronization in chaotic systems. Phys. Rev. Lett. 64, 821–824 (1990)
8. Lu, L., Zhang, C., Guo, Z.A.: Synchronization between two different chaotic systems with nonlinear feedback control. Chinese Physics 16(6), 1603–1607 (2007)
9. Sundarapandian, V.: Hybrid chaos synchronization of hyperchaotic Liu and hyperchaotic Chen systems by active nonlinear control. Internat. J. Computer Sci. Eng. Inform. Tech. 1(2), 1–14 (2011)
10. Liao, T.L., Tsai, S.H.: Adaptive synchronization of chaotic systems and its applications to secure communications. Chaos, Solitons & Fractals 11, 1387–1396 (2000)
11. Sundarapandian, V.: Adaptive synchronization of hyperchaotic Lorenz and hyperchaotic Lü systems. Internat. J. Instrument. Control Sys. 1(1), 1–18 (2011)
12. Sundarapandian, V.: Adaptive stabilization and synchronization of Lü-like chaotic attractor. Internat. J. Comp. Sci. Eng. Informat. Tech. 1(4), 15–26 (2011)
13. Yang, T., Chua, L.O.: Control of chaos using sampled-data feedback control. Internat. J. Bifur. Chaos 9, 215–219 (1999)
14. Yu, Y.G., Zhang, S.C.: Adaptive backstepping synchronization of uncertain chaotic systems. Chaos, Solitons & Fractals 27, 1369–1375 (2006)
15. Konishi, K., Hirai, M., Kokame, H.: Sliding mode control for a class of chaotic systems. Phys. Lett. A 245, 511–517 (1998)
16. Sundarapandian, V.: Global chaos synchronization of Pehlivan systems by sliding mode control. Internat. J. Comp. Sci. Eng. 3(5), 2163–2169 (2011)
17. Zhou, P., Cao, Y.X., Cheng, X.F.: A new hyperchaos system and its circuit simulation by EWB. Chinese Physics B 18(4), 1394–1398 (2009)
18. Hahn, W.: The Stability of Motion. Springer, New York (1967)

Secured Ontology Matching Using Graph Matching

K. Manjula Shenoy[1], K.C. Shet[2], and U. Dinesh Acharya[1]

[1] Department of Computer Science and Engineering, MIT, Manipal University, Manipal
{manju.shenoy,dinesh.acharya}@manipal.edu
[2] Department of Computer Engineering, NITK, Suratkal
kcshet@rediffmail.com

Abstract. Today's market evolution and high volatility of business requirements put an increasing emphasis on the ability for systems to accommodate the changes required by new organizational needs while maintaining security objectives satisfiability. This is all the more true in case of collaboration and interoperability between different organizations and thus between their information systems. Ontology mapping has been used for interoperability and several mapping systems have evolved to support the same. Usual solutions do not take care of security. That is almost all systems do a mapping of ontologies which are unsecured. We have developed a system for mapping secured ontologies using graph similarity concept. Here we give no importance to the strings that describe ontology concepts, properties etc. Because these strings may be encrypted in the secured ontology. Instead we use the pure graphical structure to determine mapping between various concepts of given two secured ontologies. The paper also gives the measure of accuracy of experiment in a tabular form in terms of precision, recall and F-measure.

1 Introduction

Researchers have developed several tools that enable organizations to share information, largely, most of these have not taken into the account the necessity of maintaining privacy and confidentiality of data and metadata of the organizations who want to share information. Consider the scenario of two different country military wanting to share information about a mission at hand while preserving the privacy of their systems. To the best of our knowledge current systems do not allow this type of information sharing.

Need for secured information sharing also exists for intra organizational information sharing too. Within the organizations different departments may use different systems which are autonomously constructed. The secure interoperability may be required here too.

Privacy should be maintained for both data and metadata. Metadata describes how data is organized (data schema), how access are controlled in the organization(the internal access control policy and role hierarchies) and the semantics of the data used in the organization(ontology).

Organizations looking to interoperate are largely using metadata like ontologies to capture the semantics of the terms used in the information sources maintained by the

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 11–18.
springerlink.com

organizations. Normally it has been assumed that these ontologies will be published by the organizations. Published ontologies from different organizations are mapped and matching rules are generated. Queries to information sources are rewritten using these matching rules so that vocabulary used in the query matches with the vocabulary of information source.

Unlike the traditional way some organizations may not like to publish their metadata or share it with other external users. Yet they want interoperation. In this case the privacy of the metadata must be preserved. The external user should not have access to ontologies in clear text. So ontologies may be encrypted and then published. The mapping system should now be able to recognize mapping in this encrypted ontology. Here we present one such system.

2 Related Work

The present ontology mapping systems can be classified into the following categories.

1. Word Similarity based: Here matching is performed based on similarity of words describing concepts, properties or names of concepts and properties occurring in the ontology.[4]
2. Structure based: Here structure of ontologies has been used for matching concepts.[5][6][7].
3. Instance based: These take the instances under concepts to find matching.[8]. These methods are further subdivided into Opaque and pattern based. In Opaque instance matching we use statistical properties like distribution, entropy and mutual information etc. In Pattern based method instance pattern are matched.
4. Inference Based: The semantics of concepts under ontologies are expressed as rules in a logical language and then the matching is performed using an inference engine.

There are also hybrid algorithms for matching ontologies.

[1] discusses need for secured data sharing in or among organization and [2] explains need for secured data mining. [3] proposes two methods for privacy preserving ontology matching. One of which is semi-automatic. And the other requires the dictionaries or thesauri or corpuses to be encrypted. Our method falls purely under structure based ontology matching which can be applied to encrypted ontologies. [4] defines a graph matching technique we used, in the literature.

3 Graph Matching Technique Used

3.1 Generalizing Hubs and Authorities[17]

Efficient web search engines such as Google are often based on the idea of characterizing the most important vertices in a graph representing the connections or links between pages on the web. One such method, proposed by Kleinberg [16], identifies in a set of pages relevant to a query search the subset of pages that are good *hubs* or the subset of pages that are good *authorities*. For example, for the query “university,” the

home-pages of Oxford, Harvard, and other universities are good authorities, whereas web-pages that point to these home-pages are good hubs. Good hubs are pages that point to good authorities, and good authorities are pages that are pointed to by good hubs. From these implicit relations, Kleinberg derives an iterative method that assigns an "authority score" and a "hub score" to every vertex of a given graph. These scores can be obtained as the limit of a converging iterative process, which is described in section below.

Let $G = (V,E)$ be a graph with vertex set V and with edge set E and let hj and aj be the hub and authority scores of vertex j. We let these scores be initialized by some positive values and then update them simultaneously for all vertices according to the following *mutually reinforcing relation*: the hub score of vertex j is set equal to the sum of the authority scores of all vertices pointed to by j, and, similarly, the authority score of vertex j is set equal to the sum of the hub scores of all vertices pointing to j:

$$\begin{cases} h_j & \leftarrow & \sum_{i:(j,i)\in E} a_i, \\ a_j & \leftarrow & \sum_{i:(i,j)\in E} h_i. \end{cases}$$

Let B be the matrix whose entry (i, j) is equal to the number of edges between the vertices i and j in G (the *adjacency matrix* of G), and let h and a be the vectors of hub and authority scores. The above updating equations then take the simple form

$$\begin{bmatrix} h \\ a \end{bmatrix}_{k+1} = \begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} h \\ a \end{bmatrix}_k, \qquad k = 0, 1, \ldots,$$

which we denote in compact form by

$$x_{k+1} = M\, x_k, \qquad k = 0, 1, \ldots,$$

Where

$$x_k = \begin{bmatrix} h \\ a \end{bmatrix}_k, \quad M = \begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix}.$$

Notice that the matrix M is symmetric and nonnegative. We are interested only in the relative scores and we will therefore consider the *normalized* vector sequence

$$z_0 = x_0 > 0, \quad z_{k+1} = \frac{M z_k}{\|M z_k\|_2}, \quad k = 0, 1, \ldots,$$

Where$\|..\|_2$ is the Euclidean vector norm. Notice that the above matrix M has the property that

$$M^2 = \begin{bmatrix} BB^T & 0 \\ 0 & B^T B \end{bmatrix},$$

and from this equality it follows that, if the dominant invariant subspaces associated with BB^T and B^TB have dimension 1, then the normalized hub and authority scores are simply given by the normalized dominant eigenvectors of BB^T and B^TB. This is the definition used in [16] for the authority and hub scores of the vertices of G. The

arbitrary choice of $z0 = \mathbf{1}$ made in [16] is shown here to have an extrenal norm justification. Notice that when the invariant subspace has dimension 1, then there is nothing particular about the starting vector **1**, since any other positive vector $z0$ would give the same result. We now generalize this construction. The authority score of vertex j of G can be thought of as a similarity score between vertex j of G and vertex *authority* of the graph

hub→authority

and, similarly, the hub score of vertex j of G can be seen as a similarity score between vertex j and vertex *hub*. The mutually reinforcing updating iteration used above can be generalized to graphs that are different from the hub–authority structure graph.

The idea of this generalization is easier to grasp with an example; we illustrate it first on the path graph with three vertices and then provide a definition for arbitrary graphs. Let G be a graph with edge set E and adjacency matrix B and consider the *structure graph*

1→2→3

With each vertex j of G we now associate three scores $xi1$, $xi2$, and $xi3$, one for each vertex of the structure graph. We initialize these scores with some positive value and then update them according to the following mutually reinforcing relation:

$$\begin{cases} x_{i1} \leftarrow & \sum_{j:(i,j)\in E} x_{i2}, \\ x_{i2} \leftarrow & \sum_{j:(j,i)\in E} x_{i1} \quad + \sum_{j:(i,j)\in E} x_{i3}, \\ x_{i3} \leftarrow & \sum_{j:(j,i)\in E} x_{i2}, \end{cases}$$

or, in matrix form (we denote by $\mathbf{x}j$ the column vector with entries xij),

$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \end{bmatrix}_{k+1} = \begin{bmatrix} 0 & B & 0 \\ B^T & 0 & B \\ 0 & B^T & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \end{bmatrix}_k, \qquad k = 0, 1, \ldots,$$

which we again denote $xk+1 = Mxk$. The situation is now identical to that of the previous example and all convergence arguments given there apply here as well. We now come to a description of the general case. Assume that we have two directed graphs GA and GB with nA and nB vertices and edge sets EA and EB. We think of GA as a structure graph that plays the role of the graphs *hub* $\longrightarrow$ *authority* and $1 \longrightarrow 2 \longrightarrow 3$ in the above examples. We consider real scores xij for $i = 1, \ldots, nB$ and $j = 1, \ldots, nA$ and simultaneously update all scores according to the following updating equations:

$$x_{ij} \leftarrow \sum_{r:(r,i)\in E_B,\ s:(s,j)\in E_A} x_{rs} + \sum_{r:(i,r)\in E_B,\ s:(j,s)\in E_A} x_{rs}.$$

This equation can be given an interpretation in terms of the product graph of GA and GB. The *product graph* of GA and GB is a graph that has $nA.nB$ vertices and that has an edge between vertices $(i1, j1)$ and $(i2, j2)$ if there is an edge between $i1$ and $i2$ in GA and there is an edge between $j1$ and $j2$ in GB. The above updating equation is then equivalent to replacing the scores of all vertices of the product graph by the sum of

the scores of the vertices linked by an outgoing or incoming edge. Equation can also be written in more compact matrix form. Let *Xk* be the *nB* × *nA* matrix of entries *xij* at iteration *k*. Then the updating equations take the simple form

$$X_{k+1} = BX_kA^t + B^tX_kA, \qquad k = 0,1,\ldots,$$

where *A* and *B* are the adjacency matrices of *GA* and *GB*. This equation is further revised by Laure Ninove [18] as follows Where X_K is replaced by S_k

$$\frac{BS_kA^t + B^tS_kA}{\| BS_kA^t + B^tS_kA\|}$$

4 Secured Ontology Mapping Using Graph Matching

First we explain the graph matching technique we used. Consider the two graphs Ga and Gb shown in Figure 1. Suppose we want to match vertex 1 of Ga with vertex 4 of Gb , we need to find how much similar the vertices 2 of Ga and 2 of Gb , and 2 of Ga and 1 of Gb.

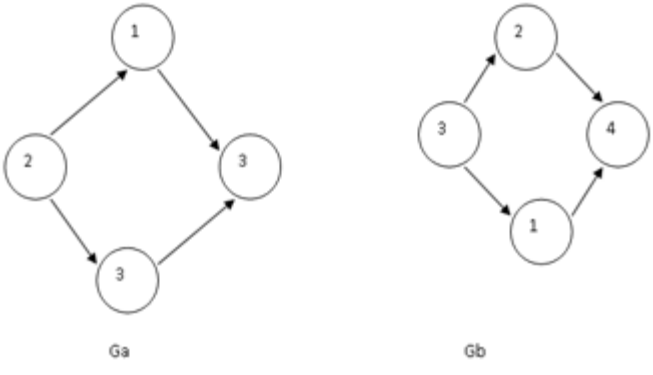

Fig. 1. Graphs to be matched

If A is the adjacency matrix of Ga and B is the adjacency matrix of Gb and S is the similarity matrix defined as follows between vertices we can get the total similarity matrix between individual vertices can be calculated using the formula

$$\frac{BSA^t + B^tSA}{\| BSA^t + B^tSA\|}$$

Here A^t stands for transpose of A. S is the initial similarity matrix. The size of S is nXm.

Where m is number of concepts in first ontology and n is number of ontology concepts in second.

The secured mapping method generates adjacency matrices based on hierarchical relationship of concepts of the encrypted ontologies as per the following algorithm. S is the unity matrix initially.

Algorithm 1. Generating Adjacency matrix for the encrypted ontology given

Let O be the ontology given and A for adjacency matrix. If n is the number of concepts in ontology O then A has order nXn.

```
1. Initialize A [i][j]=0 for all i and j between 0 and n.
2. For i= 1to n
        Begin
            Str=get i^th concept of O
            Collection = get all super classes of Str.
            For each Object x  in the  Collection
                Begin
                        For  j = 1 to n
                        If jth concept of O matches with x then
                A[i][j]=1;
                 End
             End
```

5 Results

The evaluation of the proposed system above is carried out for OAEI systematic benchmark suite. Since we compare for equality of names, and give importance to structure we need not encrypt the ontology for study of evaluation measures. The evaluation measures we considered are Precision, Recall and F-measure. Precision gives the ratio of correctly found correspondences over the total number of returned correspondences. If R is the reference alignment and A is the found alignment then the ratio for precision is

$$P(A, R) = \frac{|R \cap A|}{|A|}.$$

Recall is the ratio of correctly found correspondences to the total number of expected correspondences. The formula is

$$R(A, R) = \frac{|R \cap A|}{|R|}.$$

The following formula is used for finding F-measure.

$$M_\alpha(A, R) = \frac{P(A, R) \times R(A, R)}{(1 - \alpha) \times P(A, R) + \alpha \times R(A, R)}.$$

Here α is between 0 and 1. If α is 1 F-measure is same as precision otherwise if it is 0 then F-measure is same as recall. Usually it is taken as 0.5.

Table 1 gives the dataset and the results of experiments in terms of evaluation measures stated above.

Table 1. Result analysis

Benchmark test no	Precision	Recall	F-measure
101	1	0.8	0.88
103	1	0.8	0.88
104	1	0.8	0.88

6 Conclusion

Maintaining privacy in interoperation systems is becoming increasingly important. Ontology matching is the primary means of resolving semantic heterogeneity. Ontology matching helps establish semantic correspondence rules that are used for query rewriting and translation in interoperation systems. For information systems that want maximum privacy, the privacy of their ontologies must be maintained. Our system gives a method to map ontologies which are secured. Limitations of our system is accuracy in terms of precision and recall. The future work would be to improve this.

References

[1] Clifton, C., Doan, A., Kantarcioglu, M., Schadow, G., Vaidya, J., Elma Garmid, A., Suciu, D.: Privacy preserving data integration and sharing. In: Proc. DMKD 2004 (2004)
[2] Agrawal, R., Srikant, R.: Privacy preserving data mining. In: Proc. SIGMOD 2000 (2000)
[3] Mitra, P., Liu, P., Pan, C.C.: Privacy preserving ontology matching. In: Proc. AAAI Workshop (2005)
[4] Li, J.: LOM:Lexicon based ontology mapping tool. In: Proc. PerMIS 2004 (2004)
[5] Noy, N.F., Musen, M.A.: Anchor-Prompt: Using non local context for semantic matching. In: Proc. IJCAI 2001 (2001)
[6] Noy, N.F., Musen, M.A.: The Prompt Suite: Interactive tools for ontology mapping and merging. International Journal of Human Computer Studies 59(6) (2003)
[7] Melnik, S., Molina, H.G., Rahm, E.: Similarity flooding a versatile graph matching algorithm and its application to schema matching. In: Proc. ICDE 2002 (2002)
[8] Doan, A., Madhavan, J., Domingos, P., Halevy, A.: Learning to map be tween ontologies on the semantic web. In: Proc. WWW 2002 (2002)
[9] Melnik, S., Molina, H.G., Rahm, E.: Similarity flooding a versatile graph matching algorithm. In: Proc. ICDE 2007 (2007)
[10] Choi, N., Song, I.Y., Han, H.: A survey on ontology mapping. In: SIGMOD RECORD 2006 (2006)
[11] Shvaiko, P., Euzenat, J.: Ten Challenges for ontology matching. In: Proc. ICODAS 2008 (2008)
[12] Kalfoglou, Y., Schorelmmer, M.: Ontology mapping:The State of the Art. The Knowledge Engineering Review 18(1) (2003)
[13] Ehrig, M., Staab, S.: QOM: Quick Ontology mapping. GI Jahrestagung (1) (2004)

[14] Rahm, E., Bernstein, P.: A survey of approaches to automatic schema matching. The VLDB Journal 10(4) (2001)
[15] Shvaiko, P., Euzenat, J.: A Survey of Schema-Based Matching Approaches. In: Spaccapietra, S. (ed.) Journal on Data Semantics IV. LNCS, vol. 3730, pp. 146–171. Springer, Heidelberg (2005)
[16] Kleinberg, J.M.: Authoritive sources in a hyper linked environment. Journal of ACM (1999)
[17] Van Dooren, B., et al.: A measure of similarity between graph vertices: Application to synonym extraction and we searching. SIAM Review (2004)
[18] Ninove, L., et al.: Graph similarity algorithms. Seminar Presented at Department of Mathematical Engineering. University of Catholique de Louvain (2007)

Image Restoration Using Knowledge from the Image

S. Padmavathi, K.P. Soman, and R. Aarthi

Amrita School of Engineering, Coimbatore, Tamil Nadu, India
{s_padmavathi,r_aarthi}@cb.amrita.edu, kp_soman@amrita.edu

Abstract. There are various real world situations where, a portion of the image is lost or damaged which needs an image restoration. A Prior knowledge of the image may not be available for restoring the image, which demands for a knowledge derivation from the image itself. Restoring the lost portions of the image based on the knowledge obtained from the image area surrounding the lost area is called as Digital Image Inpainting. The information content in the lost area could contain structural information like edges or textural information like repeating patterns. This knowledge is derived from the boundary area surrounding the lost area. Based on this, the lost area is restored by looking at similar information in the same image. Experimentation have been done on various images and observed that the algorithm restores the image in a visually plausible way.

1 Introduction

The restoration process can be viewed as an algorithm that fills the gap with the information obtained from the rest of the image. The results would look natural enough that observer without prior knowledge of the original image will not notice the gaps. A Preliminary version of this is "Cloning Brush tool" in Adobe Photoshop where the user has to provide the information of what to fill in. There are two major methods of restoring the missing data. The traditional method concentrates on the structural information called the isophotes [3], which are lines of similar color. A series of differential equations are used iteratively to extend these lines in to the missing area using the information obtained from the boundary pixels. Another section of algorithms are based on Texture synthesis [1]. This method restores the missing data from an initial seed. Before a pixel is synthesized, its neighbors are sampled. Then the whole image is queried to find out a source pixel with similar neighbors. At this point, the source pixel is copied to the pixel to be synthesized. This is called as pixel based texture synthesis. Alternatively Patch based texture synthesis[6,11] could be used for increased computational efficiency, where a small area (patch) is synthesized rather than a single pixel.

Both structural and texture synthesis methods have their strengths and weaknesses. The former extends the linear structures well but introduces artifacts such as blur. The later avoids the blur but does not extend the linear structures. In general, if the thickness of the area to be restored is very less than the structural methods are advantageous. On the other hand if the thickness is large the texture methods give better results. This paper uses an algorithm similar to exemplar based method as discussed in [6]. It propagates both the texture and the structure and hence restores the missing data with a good visual quality.

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 19–25.
springerlink.com

The algorithm first fills the information in lost area that lies near the structural area and the boundary and then proceeds to other areas. The image texture is propagated by direct sampling of the source region. For such restoration the whole image is searched for an area that closely matches the boundary. The restoration is done from the boundary towards the center. This kind of filling from the boundary towards the center is called as "fill front". Moreover the filling is done in a patch-wise manner. So every time when a patch is filled, the boundary changes and the boundary is detected again and updated. To ensure a better quality of the image the algorithm checks the boundary area for sharp changes like edges and assigns more weights to the unknown pixels closer to the edges. The algorithm also gives more weights to the pixels near the boundary. The pixel with highest weight is considered for filling first. The algorithm stops when all pixels in the damaged area are restored or synthesized. The results have been obtained for various parameters.

2 Restoration Algorithm

The inpainting algorithm discussed in this paper accepts a damaged image as input with the damaged areas marked in special color. The algorithm first groups the pixels with the special color as unknown. For ease of comparison, we adopt notation similar to that used in the inpainting literature. The lost area which is the region to be filled is called as the *target* region indicated by Ω as shown in the fig. 1. The boundary between the known and the unknown area or the contour is denoted by δΩ. The unaffected image area is called the source region, Φ.

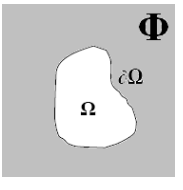

Fig. 1. Inpainting Problem

The algorithm in this paper accomplishes restoration on a patch wise manner where a patch is a small window or a square matrix of pixels. The terminologies used in this paper are derived from Crimsini et.al. For a pixel '*p*' the patch ψ_p is formed with '*p*' as the centre as shown in fig. 2a. The patch contains some unknown pixels from Ω and some known pixels from Φ with '*p*' belonging to the boundary. A patch that is similar to the known pixels is searched in the entire image (excluding the unknown pixels), (i.e.) patch that closely matches with the known pixels of ψ_p is searched in Φ. The patch that yields minimum SSD (Sum of Squared differences) value is taken as the best match. This is called as the exemplar patch. The best-match sample from the source region comes from the patch ψ_q as shown in fig. 2b. The values corresponding to the unknown pixels are copied from the best matched patch. The unknown pixels of ψ_p are copied from corresponding locations of ψ_q. A higher number of known pixels in the patch ψ_p increase the confidence of accuracy. Since the pixels are copied as

such blurring of edges is avoided. If any other patch near 'p' is considered first for restoration, edges(structure) will not be extended correctly. Hence the pixels on the boundary and the pixels near the edges have to be given a higher priority for restoration process. These priorities are termed as Confidence and Data terms respectively.

The patch priority is computed as the product of confidence term and data term. The patch with maximum priority value is restored first. Once when the highest priority patch is filled, there will be a change in the confidence values and the boundary as shown in fig. 2c. The confidence term of that particular patch will be updated as the sum of all the confidence values of the newly filled pixels divided by the total number of pixels in the patch. The new boundary is detected and the process is repeated until all the patches are filled and till the number of boundary pixels becomes zero.

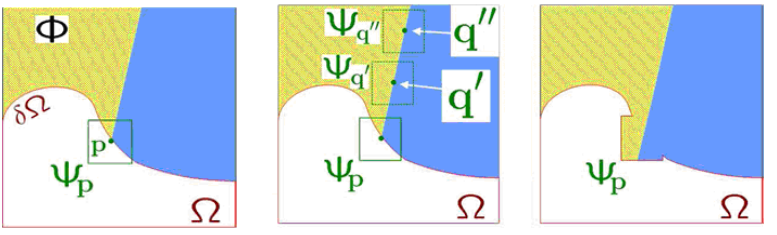

Fig. 2a. Higher Priority patch ψ_p **Fig. 2b.** Matching patches ψ_q **Fig. 2c.** ψ_p filled

Fig. 2. Inpainting of a patch ψ_p

Confidence Term C(p)

The confidence term of a patch denotes the amount of maximum reliable information about the source region in a particular patch. Initially the confidence value is assigned as 0 for the target region(Ω) and 1 for the source region(Φ). The confidence value for a patch ψ_p is calculated as in equation 1

$$C(p) = \frac{\sum_{q \in \Phi \cap \Psi_p} C(q)}{|\Psi_p|} \tag{1}$$

Where, C(q) is the confidence value of those pixels belonging to the source region and the patch ψ_p..The denominator is the area of the patch, which is the total number of pixels in the patch.

Data term D(p)

The data term gives the structure information of the image. The structure is measured in terms of magnitude of image gradient. The data term for the target region is set as 0 initially. The data term for a patch ψ_p is calculated as in equation 2

$$D(p) = \frac{\nabla \Phi_p}{\alpha} \tag{2}$$

Where $\nabla \Phi_p$ is the maximum value of the image gradient in the patch, α is a normalization factor. Sobel's Gradient operator [9] is used to calculate the image gradient.

The restoration process is given by the following algorithm:

1. Extract the manually selected target region and its initial front (boundary).
2. Repeat until there is no boundary pixel:

a. Identify the $\delta\Omega$. If the target region exit.

b. Compute priorities $P(p)$

$P(p) = C(p) * D(p)$, where C(p) is the confidence term and D(p) is the data term.

c. Find the patch Ψp with the maximum priority,

i.e., p = arg max p for all $P(p)$.

d. Find the best matching patch $\Psi q \in \Phi$ in source region that minimizes $d(\Psi p, \Psi q)$, where d is the Sum of Squared Distance(SSD).

e. Copy image data from Ψ q to Ψ p for all pixels belonging to the target region.

f. Update $C(p)$ and the D(p) for the newly filled pixels.

3 Experimentation

The algorithm is implemented in Java. Using an image manipulation software the user specifies the area to be inpainted with a special color. This image is then given to the inpainting algorithm which extracts this unique colored area as mask. The boundary of the mask and the confidence term for all pixels in the mask area are calculated. While inpainting a colored image, spurious colors are generated by the algorithm if different colored channels are inpainted separately. In order to avoid this, the data term is calculated individually for each channels and its average is considered for the total priority. The patch with highest priority is chosen for filling. For any patch in the source region the distance is calculated for each color channel and a sum of the three is stored for that corresponding patch. The patch with a minimum total distance is considered as best match. The algorithm proceeds as specified in the previous section.

For any image the patch size should be slightly greater than the smallest texture element (Texel). Hence, size of the patch is a constraint of acute interest. Change in patch size can bring about visible changes in the output of the system and its performance. The experimentation is done for different patch sizes like 3,5,7,9. The patch size once selected would be maintained through the whole process. While searching for the best samples in the source region, patches are considered in two different format: overlapping (O) and non- overlapping (NO). Experimentation is done for both cases Fig. 3 is a snapshot of the implementation showing the image to be inpainted. The user is allowed to choose the image, the patch size and the type of patches. The output of the modules of the algorithm could also be viewed. Fig 4 shows a snapshot where the image has been restored partially.

The performance of the algorithm is measured by the speed and the accuracy of restoration. The accuracy of the restoration is a usually a subjective process where an observer looks at the inpainted image and able to locate where it is modified. The objective measure can be used when a reference image is available so that the area to be inpainted is already known. When the area is not known the image is restored using Adobe Photoshop where the user specifies what has to be filled in the inpainting area and this is taken as the Reference image. In either case the accuracy and the error in reconstruction is calculated with respect to the reference image.

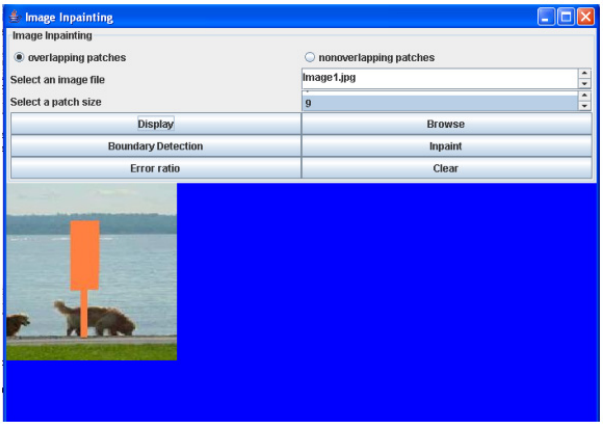

Fig. 3. Screen shot of the implementation showing the image to be inpainted

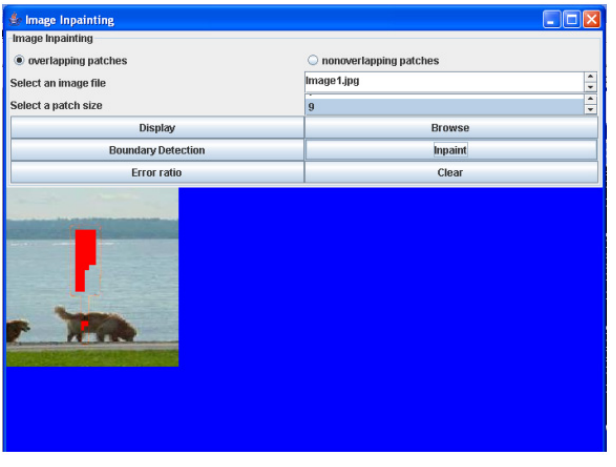

Fig. 4. Screen shot of the implementation showing a partially restored image

Fig. 5a. Before inpainting

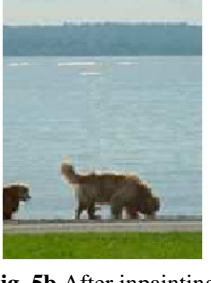

Fig. 5b. After inpainting

Fig. 5. Inpainting larger area

Experimentation is done for various sizes of target area. Fig. 5a and 6a shows a set of input images with larger (L) and smaller (S) target size respectively. Fig. 5b and 6b shows the corresponding restored images. The time and accuracy for these images are tabulated in Table 1 for various patch sizes and patch types.

Fig. 6a. Before inpainting

Fig. 6b. After inpainting

Fig. 6. Inpainting smaller area

Table 1. Performance of Restoration algorithm for images in Fig 5 and 6

Image	Nature & Size	Patch size	Color Image		Error Measure (in %)
			Time	Accuracy (in %)	
Image 1	NO/L	9	1m 45s	96.21	3.78
Image1	NO/L	7	2m 35s	96.08	3.91
Image1	NO/L	5	5m 8s	96.01	3.98
Image1	NO/L	3	7m 8s	95.83	4.17
Image1	O/L	9	1m 50s	97.08	2.92
Image1	O/L	7	2m 41s	96.68	3.32
Image1	O/L	5	5m 50s	96.05	3.95
Image1	O/L	3	7m 17s	95.40	4.60
Image1	NO/S	9	1m 30s	95.04	4.96
Image1	NO/S	7	2m 9s	96.60	3.40
Image1	NO/S	5	4m 59s	96.76	3.24
Image1	NO/S	3	6m 53s	97.42	2.58
Image1	O/S	9	1m 43 s	95.67	4.33
Image1	O/S	7	2m 32s	96.16	3.84
Image1	O/S	5	4m 23s	96.87	3.13
Image1	O/S	3	6 m 30s	97.02	2.98

In Table 1 NO and O represents non overlapping patches and Overlapping patches respectively. L and S represents larger and smaller target area respectively. It could be observed that the overlapping patches gives better accuracy and lesser error than Non-Overlapping patches which implies that reconstruction is better in the former case. The time consumed for reconstruction is more for Larger and Overlapping patches when compared to Smaller and Non-Overlapping Patches.

4 Conclusion

The experiments have been conducted on various images involving natural scenes, regular structured images, different shapes and sizes of the area to be restored. The algorithm is robust towards changes in shape and topology of the region to be restored. It preserves edge sharpness and avoids spurious colors. It is observed from the results that, overlapping small patches gives better results even for a composite textured image such as a natural scene image. But when the image has uniform texture or no structure, non overlapping patches of any size gives good results. The system gives moderate result when the information for filling is not available anywhere in the rest of the image.

References

[1] Ashikhmin, M.: Synthesizing natural textures. In: Proc. ACM Symposium on Interactive 3D Graphics, pp. 217–226. Research Triangle Park, NC (2001)

[2] Chan, T.F., Shen, J.: Non-texture inpainting by curvature-driven diffusions (CDD). Journal of Visual Communication and Image Representation 4(12), 436–449 (2001)

[3] Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: Proc. ACM Conference Computer Graphics, SIGGRAPH, New Orleans, LU, pp. 417–424 (July 2000), http://mountains.ece.umn.edu/

[4] de Bonet, J.S.: Multi resolution sampling procedure for analysis and synthesis of texture images. In: Proc. ACM Conference Computer Graphics, SIGGRAPH, vol. 31, pp. 361–368 (1997)

[5] Harrison, P.: A non-hierarchical procedure for re-synthesis of complex texture. In: Proc. Int. Conf. Central Europe Computer Graphics, Visualization and Computer Vision, Plzen, Czech Republic (February 2001)

[6] Criminisi, A., Pérez, P., Toyama, K.: Region Filling and Object Removal by Exemplar-Based Image Inpainting, Microsoft Research, Cambridge (UK) and Redmond (US)

[7] Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: Proc. ACM Conference Computer Graphics, SIGGRAPH, New Orleans, LU, pp. 417–424 (July 2000)

[8] Bertalmio, M., Vese, L., Sapiro, G., Osher, S.: Simultaneous structure and texture image inpainting. In: Proc. Conference Computer Vision Pattern Recognition, Madison, WI (2003)

[9] Gonzalez, R.C., Woods, R.E.: Digital Image Processing

[10] Rutherford, H.: A Practial Introduction to Image Processing using Java. Pearson University

[11] Hertzmann, A., Jacobs, C., Oliver, N., Curless, B., Salesin, D.: Image analogies. In: Proc. ACM Conf. Comp. Graphics, SIGGRAPH, Eugene Fiume (August 2001)

Harmony-Based Feature Weighting to Improve the Nearest Neighbor Classification

Ali Adeli[1,2], Mehrnoosh Sinaee[1], M. Javad Zomorodian[1,3], and Ali Hamzeh[1]

[1] Department of Computer Science & Engineering, Shiraz University, Shiraz, Iran
[2] Institute of Computer Science, Bojnurd Darolfonoun Technical College, Bojnurd, Iran
[3] Institute of Computer Science, Shiraz Bahonar Technical College, Shiraz, Iran
{aliadeli,mehrnooshsinaee,jzomorodian}@shirazu.ac.ir,
ali@cse.shirazu.ac.ir

Abstract. This paper introduces the use of Harmony Search with novel fitness function in order to assign higher weights to informative features while noisy irrelevant features are given low weights. The fitness function is based on the Area Under the receiver operating characteristics Curve (AUC). The aim of this feature weighting is to improve the performance of the *k*-NN algorithm. Experimental results show that the proposed method can improve the classification performance of the *k*-NN algorithm in comparison with the other important method in realm of feature weighting such as Mutual Information, Genetic Algorithm, Tabu Search and chi-squared ($\chi 2$). Furthermore, on synthetic data sets, this method is able to allocate very low weight to the noisy irrelevant features which may be considered as the eliminated features from the data set.

Keywords: AUC, Harmony Search, Feature weighting, Noisy feature elimination, *k*-NN.

1 Introduction

In non-parametric density estimation algorithms, the distribution of data is calculated without any particular assumption on its parameters. Two popular approaches in these algorithms are: Kernel Density Estimation (KDE) and k-Nearest Neighbor (*k*-NN). *K*-NN is a simple classifier that has been used in various real world applications. In some cases *k*-NN is vulnerable with some problems, such as few instances, noisy data and too many features, which decrease the performance of *k*-NN. To improve the performance of this classifier, many solutions are introduced. One of the approaches to solve those mentioned problems is searching in feature space to find optimal subset of features that can improve the classification accuracy of *k*-NN. This goal can be achieved by assigning the weight to all features in order to eliminate irrelevant ones from noisy data sets. In this paper, we attack this problem and introduce a novel algorithm to deal with. At first, proposed approach assigns different weight to feature using Harmony Search and AUC measure as the fitness function. In other words, Harmony Search with statistical measure as a fitness function has been used to allocate optimal weights to all features to achieve better classification accuracy.

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 27–36.
springerlink.com

This paper is organized as follows: in section 2, related work in this domain is reviewed. AUC and Receiver Operating Characteristics (ROC) curve are described in section 3. The proposed method is presented in section 4. Section 5 includes data set, experimental results and discussion, and conclusion is mentioned in the last section of paper.

2 Related Works

There is much research considering the problem of feature selection. It has been used when I) the number of features is larger than the number of training data, II) the number of features is too large for feasible computation III) many features include noisy value. These issues can result in a significant drop in classification performance.

Many methods have been proposed for feature ranking, i.e. Weight Adjusted k Nearest Neighbor (WAKNN) which is introduced by Han [7] to overcome the problem of curse of dimensionality. He implemented his idea on the text classification using *k*-NN. In his work, each attribute takes a weight using the Mutual Information (MI) between each word and the class variable. In the domain of feature weighting, another work refers to Weighted Artificial Immune Recognition System (WAIRS) [10]. In this paper, MI is the main algorithm for feature weighting. Note that the weighted attributes were added to the AIRS. Classification is the final step of AIRS algorithm that is performed by *k*-NN.

Jankowski and Copernicus recommended weighted *k* Nearest Neighbor (WkNN) idea [8]. In each fold of their algorithm, the initial weights for all features are set to 1. During each fold, the values of the weights are summed (subtracted) with Δ value. If the upsdated value can improve the accuracy of the *k*-NN, the new value is replaced with old one for corresponding feature. After each fold, weighting procedure returns a vector of weights. After all folds, i.e. 10 folds, the algorithm computes a normalized vector which is a summation of 10 vectors.

GAW is a common solution for weighting attributes that is suggested by Tang and Tseng [11]. GAW is based on the Genetic Algorithm (GA) with real representation. In this paper, weighing approach is used to improve the accuracy of Weighted Fuzzy *k*-NN (WFKNN) classifier. Guvenir and Akkus studied on Weighted k Nearest Neighbor Feature Projection (WkNNFP) [6]. In WkNNFP, Single Feature Accuracy (SFA) procedure is utilized for feature weighting. In SFA, weight of each feature is determined according to accuracy which is obtained by considering only this feature.

Tabu Search (TS) is proposed as a weighting method in [13]. In this paper, a Hybrid Tabu Search/*K*-NN algorithm is proposed to perform both feature selection and feature weighting simultaneously. In other words, *k*-NN is used each weight set generated by TS. It searches heuristically in a local neighborhood area and moves from a solution to its best admissible neighbor.

The proposed chi-squared ($\chi 2$) Feature Weighting ($\chi 2$FW) method can be classified as a mutual information approach for assigning features weights [12]. In this sense, the mutual information (the Chi-Squared statistical score) between the values of a feature and the class of the training instances are used to assign feature weights [12].

The algorithm uses Sequential Weighting as the weighting criteria in order to give weights to the features. The weighting criteria ranks features according to their $\chi 2$ scores. In other words, the features having the lowest $\chi 2$ score have their weights set to 1, those with the second lowest-scored features have their weight set to 2 and so on. The process goes on until weights are assigned to the highest $\chi 2$ scored features [12]. In the wide range of weighting approach, algorithm processes the usefulness of each feature independently. So, the non-linear interaction between features has been ignored. While in the proposed method, each Harmony vector takes into account this interaction and also their importance (not its importance) on the classification problem.

3 Receiver Operating Characteristic (ROC)

The Receiver Operating Characteristic (ROC) curve is a two dimensional illustration of the classifier performance. It is suitable solution to analyze the classification accuracy in the binary class problem. For this purpose, the ROC curve plots the True Positive rate (Sensitivity) versus False Positive rate (1- Specificity). The ROC is a strong and statistical tool to compare binary classifiers. Sensitivity and Specificity are described in (1) and (2). To plot the ROC curve, the sensitivity and specificity need to be calculated as follows [3]:

- True Positive (TP) = number of predicted positive cases that are actually positive.
- True Negative (TN) = number of predicted negative cases that are actually negative.
- False Positive (FP) = number of predicted positive cases that are actually negative.
- False Negative (FN) = number of predicted negative cases that are actually positive.

$$Sensitivity = \frac{TP}{TP+FP} \tag{1}$$

$$Specifity = \frac{TN}{TN+FP} \tag{2}$$

The AUC is a part of the area of the unit square. The AUC is a scalar value, in interval [0--1], to show the discriminative power of binary classifiers. If the value of AUC is less than "0.5", it shows undesirable result, but if the AUC value of classifier is close to "1", it shows a remarkable performance for binary classification. Equation (3) shows the AUC formula.

$$AUC = \sum_{k=1}^{n}(X_k - X_{k-1})(Y_k - Y_{k-1}) \tag{3}$$

3 Proposed Method

The aim of this study is to improve the classification accuracy of the *k*-NN algorithm. One of the best solutions to improve the *k*-NN classification is that the informative features are given large weights while noisy irrelevant features are given low weights.

In this respect, a great approach is needed to select the best features easily. For this purpose, first contribution of the paper has focused on the Harmony-based feature weighting. The Harmony Search is one of the best and popular search tools. Harmony Search as a weighting procedure is a novel approach which can classify input instances with informative and relevant feature. The second contribution of the paper is to use the AUC as a fitness function for Harmony vectors.

First of all, the data set has been split to the unseen data and training sets. Next, the 10-fold cross validation function has been used to validate the *k*-NN. Then in each fold, the Harmony Search procedure is called. In Harmony algorithm, a population of n Harmony vector has been produced randomly and the fitness function of population (Harmony Vectors) is computed. Note that Harmony vectors includes real value in range [0--1]. After that, fitness value of each Harmony vector is calculated using AUC function. Then evaluated Harmony vectors are used in evolutionary progress. The cycle of Harmony Search will be described later. The evolutionary process has continued until the conditions are satisfied, i.e. variance of fitness value for the best Harmony vector is lower than a predefined threshold. After each fold, the training error should be computed with the validation set. For this propose, Harmony algorithm returns a vector (in size of features) with the best real values in range [0--1] (each value refers to corresponding feature). After that, the weighted features are stored to the *k*-NN algorithm for classification. Note that all weights will be employed in edited version of Minkowski metric (8) to compute the distance between training and testing instances. After 10 folds, best features are given higher weights while the irrelevant ones are given low weights and then they have been used in the *k*-NN. Finally, the testing error has been computed.

3.1 Harmony Search (HS)

Harmony Search (HS) is a metaheuristic algorithm that is proposed by Geem et al. [2]. HS is inspired by improvisation process of music player and mimicking its phenomenon [2]. In improvisation process, musician plays a note to find best Harmony. Similar to this process, in engineering problem, a decision variable generates a value to find global optimum. HS is a derivative-free, does not require initialization setting for decisions values, free from divergence and can deal with both types of variables (discrete and continuous) [2,9].

A new Harmony vector can be made by choosing a pitch from following rules: 1) playing one pitch from memory (musician's memory); 2) playing a pitch near to pitches in the memory; and 3) playing a random pitch from possible range of pitches. Likewise, in HS algorithm, the value of a decision variable is selected according to one of the following rules: 1) selecting a value from Harmony Memory (HM); 2) selecting a value near to values of HM; and 3) selecting a random value in the possible range of values. HS includes some parameters which are described as follow:

HM (Harmony Memory) encloses all the generated Harmony vectors. Equation (4) shows the memory. **HMS** (HM Size) determines the number of Harmony vectors in HM. Each vector is a solution for optimization problem. Note that in experimental results, value of HMS is set to 100.

$$HM = \begin{bmatrix} x_1^1 & x_2^1 & \dots & x_n^1 \\ x_1^2 & x_2^2 & \dots & x_n^2 \\ & & & \\ x_1^{HMS} & x_2^{HMS} & \dots & x_n^{HMS} \end{bmatrix} \tag{4}$$

HMCR (HM Considering Rate) is a probability number to select a value from HM for decision variable (rule 1 and 2) that is illustrated in (5). To satisfy the third above rule, we use (1-HMCR) to choose a random value out of the HM but in possible range of values. HMCR is a high value because in the music domain, each musician has a specific methodology and follows its method in the most melodies. Note that in experimental results, value of HMCR is set to 0.9.

$$x_i^{'} = \begin{cases} x_i^{'} \in \left\{ x_i^1, x_i^2, \dots, x_i^{HMS} \right\} & \text{with probability HMCR} \\ x_i^{'} \in X_i & \text{with probability (1-HMCR)} \end{cases} \tag{5}$$

PAR (Pitch-Adjusting Rate) is a probability number to determine the rate of small changes in values of the variable. We will see in Algorithm 1 that the probability of PAR is checked inside the condition of HMCR. In other words, if the random generated value is lower than the HMCR, the condition of PAR will be checked, otherwise the PAR condition is not checked. This probability is used to satisfy second above rule. The PAR value is calculated according to (6). The value of the PAR is small because musician follows its method and selects rarely a random Harmony. Note that in experimental results, value of PAR is set to 0.1.

$$x_i^{'} = \begin{cases} x_i^{'} \pm rand(0,1).bw & \text{with probability PAR} \\ x_i^{'} & \text{with probability (1-PAR)} \end{cases} \tag{6}$$

bw is an arbitrary distance bandwidth that shows the range of small changes in values of variable. **MaxImp** is the maximum number of iterations. Note that in experimental results, value of *bw* is set to 0.001.

3.2 Fitness Function

The fitness function used in the HS is based on the AUC [13]. In the AUC algorithm, each instance takes a probability score based on its label of neighbors. To compute the probability, first, the distance of each sample to the others has been calculated using (8) which uses the weight of all features. Then, the labels of *k* nearest neighbors to this sample are considered. The score of each sample has been computed with (7).

$$\text{Score(i)} = \frac{\text{no. of positive NN to Sample i}}{\text{k}} \tag{7}$$

The Minkowski metric for the weighted feature is changed to (8) which considers the weights of attributes. In (8), w_i is the weight of i^{th} feature.

$$L_k(a,b) = \left(\sum_{i=1}^{d} w_i \left| a_i(x) - b_i(y) \right|^k \right)^{\frac{1}{k}} \tag{8}$$

It is also referred to as the L_k norm. So, the Euclidean distance is the L_2 norm, and the L_1 norm refers to the Manhattan one. After the computation of scores, TP and FP rates are measured, and the ROC curve is plotted. At last, to compute the value of the AUC, the area of all trapezoids, which are located under the ROC curve, is calculated. The summation of all these areas can be considered as the AUC and fitness of Harmony vector.

4 Experimental Results

In the testing phase, 10-fold cross validation used to validate empirical results. After each fold, the validation set has been used to compute the training error of the *k*-NN classification. After 10 fold, the testing error of the *k*-NN classification has been calculated using the testing set (unseen data). The results of the testing have been reported in Tables 2 and 3. The data sets used for the analysis of the model have been indexed in Table 1. Experimental results were achieved in two types. In the first type, our presented method deals with 9 binary class (multi-class) distribution of data which are mentioned in Table 3. All data sets were chosen from UCI repository [1]. Next type of testing is applied in order to analyze the behavior of the proposed method on generated irrelevant features (Table 2). For the second type, our method tested on the seven synthetic data sets which are randomly generated with some relevant and irrelevant features [12, 13]. All the synthetic data sets contain 500 samples in the binary class distribution. The values of all features (relevant/irrelevant) are randomly picked from distribution in interval [0--1]. In all synthetic data sets, a data point belongs to positive class if the average value of relevant features for this instance is smaller than the threshold; otherwise it belongs to negative class. The threshold is set as the average values of all features in whole data. So, for each data set, the threshold is deterministic. Table 2 compares the performance of the proposed method with the simple *k*-NN on the synthetic data sets.

4.1 Discussion

In this section, results of the proposed method are compared with some important feature weighting methods, Mutual Information (MI), Genetic Algorithm (GA), Tabu Search (TS) and chi-squared Feature Weighting (χ2FW). Note that the basic classifier used in the all mentioned weighting methods is the *k*-NN. Experimental results show that the proposed method can improve the classification performance of the *k*-NN. Furthermore, in a number of cases, *k*-NN classifier with the Harmony-based weighting method can perform better than the simple *k*-NN (without feature ranking). In Table 2, effecting of weighting method on *k*-NN classification is presented. For this purpose, we generated some data sets with different number of relevant and irrelevant features. Experimental results show that in all cases of generated data sets, the proposed method outperforms the simple *k*-NN without weighting mechanism.

Algorithm 1 HS procedure

Input(s): Parameter setting

Output: Best Harmony solution (vector)

1: Initial new vector (x') with zero values for all features
2: **for** solution :=1 to HMS
3: Generate random Harmony vector in range [0 1]
4: call the fitness function to compute the fitness of solution
5: **end for**
6: **for** iteration :=1 to MaxImp
7: **for** feature :=1 to no. of features
8: **if** rand(0,1) < HMCR **then**
9: $x_{iteration}' \leftarrow$ select randomly a value of column *feature* from HM
10: **if** rand(0,1) < PAR **then**
12: $x_{iteration}' \leftarrow x_{iteration}' + \text{rand}(0,1).bw(\text{i})$;
13: **end if**
14: **else**
15: $x_{iteration}' \leftarrow$ randomly select any pitch within bounds
16: **end if**
17: **end for**
18: calculate the fitness of new vector x'
19: **if** fitness(x') > fitness (worst vector)
20: replace the new vector x' with the worst one
21: **end if**
22: **end for**

Table 1. Data sets is used in this experiment. Number of features and number of samples in each data set is mentioned.

Dataset	# features	# samples	# class
Glass	10	214	6
Ionosphere	34	351	2
Iris	4	150	3
Hepatitis	19	155	2
Pima	8	760	2
Sonar	60	208	2
Soybean	35	307	19
Vote	16	435	2
WBC	10	699	2

Table 2. Empirical results of classic *k*-NN and the proposed approach when irrelevant features incorporated into the synthetic data sets

# Relevant features	# Irrelevant Features	Feature weighting	Original *k*-NN
4	6	**16.21±1.24**	25.01±0.67
5	5	**22.12±1.36**	24.54±1.13
5	8	**16.48±0.86**	22.87±1.32
6	4	**24.72±1.01**	26.65±0.63
8	5	**18.54±0.58**	20.72±1.78
10	10	**22.04±0.53**	26.98±1.54

Table 3. Comparison of classification errors between the proposed method and the other weighting methods. The best results on each data set is highlighted in bold face. Note that the last column refers to results of the proposed method.

	k-NN without weighting	Weighting based				
		MI	GA	TS	$\chi 2$ FW	Proposed method
WBC	93.99	95.56	94.52	95.02	96.63	**97.82**
Glass	88.43	92.78	89.56	90.40	90.56	**94.03**
Hepatitis	84.51	81.29	**87.72**	84.25	80.74	85.58
Ionosphere	89.74	89.46	85.48	93.80	90.35	**95.72**
Iris	93.33	92.67	96.84	96.70	95.02	**97.73**
Pima	69.02	75.01	67.36	74.59	75.97	**77.43**
Sonar	85.03	95.95	89.85	94.20	92.51	**97.10**
Soybean	89.01	91.55	**92.08**	90.78	89.65	85.19
Vote	92.93	95.64	92.65	94.03	94.52	**96.13**

In Table 3, the proposed method outperforms the rest on the data sets such as WBC, Glass, Ionosphere, Iris, Pima, Sonar and Vote. In comparison with the mentioned weighting methods for *k*-NN classifier, experimental results prove that the feature weighting scheme based on Harmony Search is an impressive solution to improve the accuracy of *k*-NN classifier.

WAKNN computes the weight of each feature according to value of MI between this feature and class label [7]. The reason of WAKNN's weak performance is that they process the usefulness of each feature independently. So, the nonlinear interaction between features has been ignored. In other words, WAKNN considers

the correlation between each feature and the class label independently from other features. In some cases, there are two features which the correlation between each feature and the class label is low, but the correlation between the combination of them as a features subset and the class label is high. However in the proposed method, each Harmony vector considers the contribution of all features on the classification problem. The Table compares the proposed method with GAW which gives weight to features according to GA [11]. In our approach, one of the important advantages is AUC fitness function. In GAW, classification accuracy rate of the test set (known instances which were tested) is employed for fitness function. Comparison between fitness functions of GA and HS illustrates that the AUC is a dominant function and assign fitness to Harmony vector with high confidence because of its statistical property. So, *k*-NN classifier with the HS-based weighting algorithm and the AUC fitness function outperforms the *k*-NN with Genetic-based feature weighting.

The problem of TS is that this kind of search causes the objective function to deteriorate [13]. In other words, it may be fall into local optimum without any reaching to the best set of weights (in task of feature weighting). Chi-squared Feature Weighting (χ2FW) is a weighting method that is calculated according to (9). In (9), i and j are discrete variables which can assume l and c possible values, respectively. n_{ij} and e_{ij} are the observed frequency and the expected frequency, respectively. Similar to WAKNN, the chi-squared method processes the usefulness of each feature independently. So, the nonlinear interaction between features has been ignored. In some cases, we need to analyze the effect of a group of features on the classification problem instead of considering just one feature.

$$\chi^2 = \sum_{i=1}^{l} \sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \tag{9}$$

Our algorithm shows poor performance on the Soybean and Hepatitis data sets. The reason of poor performance is that the weights are inefficient for features.

5 Conclusion

In this paper, a novel method has been introduced for feature weighting. The proposed method of feature weighting is based on HS. The best Harmony vector returns real values in range [0--1] for all features. The weighting of features is based on the fitness of the Harmony vector. The fitness of Harmony vector has been calculated using the AUC, a statistical measure to compare classifiers. The experimental results show that the proposed method improves the *k*-NN classification. In some cases, the proposed method helps the *k*-NN to result in more accurate classification than some other method in the realm of feature weighting such as MI, TS and GA. Furthermore, in the synthetic datasets, this method is able to allocate very low weight to the noisy irrelevant features which may be considered as the eliminated features from the dataset.

References

1. Blake, L., Merz, C.J.: UCI repository of machine learning databases, http://www.ics.uci.edu/~mlearn/MLRepository.html
2. Das, S., Mukhopadhyay, A., Roy, A., Abraham, A., Panigrahi, B.K.: Exploratory power of the harmony search algorithm: analysis and improvements for global numerical optimization. IEEE Transactions on Systems Man and Cybernetics Part B 41(1), 89–106 (2011)
3. Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis, vol. 7. Wiley (1973)
4. Eiben, A.E., Smith, J.E.: Introduction to Evolutionary Computing, vol. 12. Springer (2003), http://www.mitpressjournals.org/doi/abs/10.1162/evco.2004.12.2.269
5. Fawcett, T.: Roc graphs, Notes and practical considerations for data mining researchers ROC graphs. Intelligent Enterprise 31 (HPL-2003-4) 28 (2003)
6. Guvenir, H.A., Akkus, A.: Weighted k nearest neighbor classification feature projections. In: Proc. of the Twelfth International Symposium on Computer and Information Sciences, ISCIS XII, pp. 44–51 (1997)
7. Han, E.-H(S.), Karypis, G., Kumar, V.: Text Categorization Using Weight Adjusted *k*-Nearest Neighbor Classification. In: Cheung, D., Williams, G.J., Li, Q. (eds.) PAKDD 2001. LNCS (LNAI), vol. 2035, pp. 53–65. Springer, Heidelberg (2001), http://dl.acm.org/citation.cfm?id=646419.693652
8. Jankowski, N.: Discrete feature weighting selection algorithm. In: Proceedings of the International Joint Conference on Neural Networks, vol. 1, pp. 636–641 (2003)
9. Mahdavi, M., Fesanghary, M., Damangir, E.: An improved harmony search algorithm for solving optimization problems. Applied Mathematics and Computation 188(2), 1567–1579 (2007)
10. Seeker, A., Freitas, A.: Wairs improving classification accuracy by weighting attributes in the airs classifier. In: IEEE Congress on Evolutionary Computation, CEC 2007, pp. 3759–3765 (2007)
11. Tang, P.H., Tseng, M.H.: Medical data mining using BGA and RGA for weighting of features in fuzzy k-NN classification. In: International Conference on Machine Learning and Cybernetics, vol. 5, pp. 3070–3075 (2009)
12. Vivencio, D., Hruschka, E., Nicoletti, M., dos Santos, E., Galvao, S.: Feature-weighted k-nearest neighbor classifier. In: FOCI 2007, pp. 481–486 (2007)
13. Zomorodian, M.J., Adeli, A., Sinaee, M., Hashemi, S.: Improving Nearest Neighbor Classification by Elimination of Noisy Irrelevant Features. In: Horng, M.-F. (ed.) ACIIDS 2012, Part II. LNCS, vol. 7197, pp. 11–21. Springer, Heidelberg (2012)

Effectiveness of Different Partition Based Clustering Algorithms for Estimation of Missing Values in Microarray Gene Expression Data

Shilpi Bose[1], Chandra Das[1], Abirlal Chakraborty[2], and Samiran Chattopadhyay[2]

[1] Department of Computer Science and Engineering, Netaji Subhash Engineering College, Kolkata- 700 152, India
bose.shilpi08@gmail.com, chandradas@hotmail.com
[2] Department of Information Technology, Jadavpur University, Kolkata- 700 092, India
abir126@gmail.com, samiranc@it.jusl.ac.in

Abstract. Microarray experiments normally produce data sets with multiple missing expression values, due to various experimental problems. Unfortunately, many algorithms for gene expression analysis require a complete matrix of gene expression values as input. Therefore, effective missing value estimation methods are needed to minimize the effect of incomplete data during analysis of gene expression data using these algorithms. In this paper, missing values in different microarray data sets are estimated using different partition-based clustering algorithms to emphasize the fact that clustering based methods are also useful tool for prediction of missing values. However, clustering approaches have not been yet highlighted to predict missing values in gene expression data. The estimation accuracy of different clustering methods are compared with the widely used KNNimpute and SKNNimpute methods on various microarray data sets with different rate of missing entries. The experimental results show the effectiveness of clustering based methods compared to other existing methods in terms of Root Mean Square error.

Keywords: Microarray analysis, missing value estimation, c-means, fuzzy c-means, possibilistic c-means, fuzzy possibilistic c-means.

1 Introduction

Recent advancement of microarray technologies has made the experimental study of gene expression data faster and more efficient. Microarray techniques, such as DNA chip and high density oligonucleotide chip are powerful biotechnologies as they are able to record the expression levels of thousands of genes simultaneously [1].

The data generated in a set of microarray experiments are usually gathered in a matrix with genes in rows and experimental conditions in columns. Frequently, these matrices contain missing values (MVs). This is due to the occurrence of imperfections during the microarray experiment (e.g. insufficient resolution, spotting problems, deposition of dust or scratches on the slide, hybridization failures etc.) that create

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 37–47.
springerlink.com

suspected values, which are usually thrown away and set as missing [2]. In large-scale studies involving thousands to tens of genes and dozens to hundreds of experiments, the problem of missing values may be severe. Virtually every experiment contains some missing entries and more than 90% of genes are effected. The presence of missing gene expression values constitutes a problem for downstream data analysis, since many of the methods employed, such as principal component analysis [3] or singular value decomposition [4] (e.g. classification and model-based clustering techniques) require complete matrices . Due to economic reasons or biological sample availability, repeating the microarray experiments in order to obtain a complete gene expression matrix is usually not feasible and also analysis results can be influenced by the estimation of replacing the missing values. Thus, in order to minimize the effect of missing values on analysis and avoid improper analysis, missing value estimation is an important preprocess.

Generally, the procedures for dealing with the randomly present missing data can be grouped into three categories [5], [2]: (1) Ignorance-based procedures: This is the most trivial approach to deal with data sets when the proportion of complete data is small, but the elimination brings a loss of information; (2) Model-based procedures: This is a missing data recovery method, which defines a model for the partially missing data. However, the complexity of the method prevents the applications of large data sets; (3) Imputation-based procedures: This is the type of missing data substitution methods, which fill the missing values by certain means of approximation. Statistical imputation belongs to this category, where the missing values are substituted by a statistically inspired value that has a high likelihood for the true occurrence, for example the mean values computed from the set of non-missing data records.

There are several simple ways to deal with missing values such as deleting genes with missing values from further analysis, filling the missing entires with zeroes, or imputing missing values of the average expression level for the gene (’row average’) [2] etc. Two advanced estimation methods for missing value estimation in microarray data have been proposed by Troyankaya et al. [5]; a weighted K-nearest neighbor method (KNNimpute) and a singular value decomposition method (SVDimpute). KNNimpute method is proposed as a robust and sensitive method for missing value estimation. It uses the KNN procedure to select genes, and uses weighted linear combinations to predict missing values. Recently, there is an estimation method called sequential K-nearest neighbor method (SKNNimpute) [6] for microarray data. This imputes missing values sequentially from the gene having least missing values and uses the imputed value for the latter imputation. Efficiencies of KNNimpute and SKNNimpute are better than the above mentioned simple methods in terms of missing value prediction error on non time series or noisy data. SVDimpute that takes all gene profile correlation information into consideration yields best results on time series data with low noise levels. However, estimation abilities of KNNimpute and SKNNimpute depend on the important model parameter K-value, the number of gene neighbor used to estimate the missing value. The parameter is usually specified by the user, which requires the user have some domain knowledge. There is no theoretical way, however, to determine these parameters appropriately. Several other methods have also been developed to estimate missing values. Bayesian principal component analysis (BPCA) [7] is shown to perform exceptionally well [8], [9]. However, BPCA

is a sophisticated method that is highly dependent on the number of principal axes [8]. The fixed-rank approximation algorithm (FRAA) proposed by Friedland et al. [10] carries out the estimation of all missing entries in the gene expression data matrix simultaneously based on the singular value decomposition (SVD) method. Local least-squares imputation (LLSimpute) by Kim et al. [11] exploits the local similarity structures in the data and uses the least-squares optimization method to find the missing values that are represented as a linear combination of similar genes. However, the prediction error generated using these methods still impacts on the performance of statistical and machine learning algorithms including class prediction, class discovery, and differential gene identification algorithms [12]. There is, thus, considerable potential to develop new techniques that will provide minimal prediction errors for different types of microarray data including both time and non-time series sequences.

Current research demonstrates that if the correlation/similarity between genes is exploited then missing value prediction error can be reduced significantly [13] in gene expression data. Cluster analysis [14], which partitions the given data set into distinct subgroups, is also applied to predict missing values in microarray data. Intuitively, objects in a cluster are more similar to each other than those belonging to different clusters. In this sense, objects in a cluster are more correlated with each other, whereas objects in different clusters are less correlated. As it can partitions different objects into groups, based on some similarity/dissimilarity criterion, it can also be used to discover structures based on similarity/dissimilarity in gene expression data without providing any interpretation. After clustering, missing values present in a gene can be predicted more accurately from other similar genes belonging to the same cluster.

In this paper, prediction accuracies are given for estimation of missing values in microarray gene expression data with respect to RMS error, using different partition based clustering algorithms. The effectiveness of the partition-based clustering methods, along with a comparison with SKNN and KNN imputation methods, is demonstrated on three microarray data sets.

2 Different Partition-Based Clustering Algorithms for Estimaion of Missing Values

In this section different partition-based clustering algorithms are described and then a new imputation method has been demonstrated to predict missing values in microarray gene expression data.

2.1 Notation

Throughout this paper, microarray data are represented by matrices with rows corresponding to genes and columns to experimental conditions. In particular, G represents original data matrix (with real MVs), while S is a complete gene expression matrix without any missing values with N genes and E experiments (with N >> E) after preprocessing G. In this S matrix, data are randomly deleted to create test data matrix T. X represents a set containing N number of genes. Every gene contains E number of attributes.

2.2 c-Means Clustering Algorithm

The algorithm proceeds by partitioning N number of objects into c nonempty subsets. During each partition, the centroids or means of the clusters are computed. This process iterates until the criterion function converges. Typically, the square-error criterion is used, defined as

$$E = \sum_{i=1}^{K} \sum_{x_k \in U_i} |x_k - m_i|^2 \tag{1}$$

The main steps of the c-means algorithm [15] are as follows:

1) Arbitrarily choose c number of object from X and they are assigned in m_i, i = 1 to c as initial cluster means.
2) Assign each data object x_k to the cluster U_i for the closest mean.
3) Compute new mean for each cluster using

$$m_i = \frac{\sum_{x_k \in U_i} x_k}{|U_i|} \tag{2}$$

where $|U_i|$ is the number of objects in cluster U_i.
4) Iterate until criterion function converges, i.e., there are no more new assignments.

2.3 Fuzzy c-Means (FCM) Clustering Algorithm

This is a fuzzification of the c-means clustering algorithm. It partitions a set of N objects $\{x_k\}$ into c clusters by minimizing the objective function

$$J = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{ik})^p ||x_k - m_i||^2 \tag{3}$$

where $1 \leq p < 1$ is the fuzzifier, m_i is the i^{th} cluster center, $\mu_{ik} \in [0, 1]$ is the membership of the k^{th} pattern and ||.|| is the distance norm, such that

$$m_i = \frac{\sum_{k=1}^{N} (\mu_{ik})^p x_k}{\sum_{k=1}^{N} (\mu_{ik})^p} \tag{4}$$

and

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}} \right)^{\frac{2}{p-1}}} \tag{5}$$

$\forall i, d_{ik} = ||x_k - m_i||^2$, subject to $\sum_{i=1}^{c} \mu_{ik} = 1, \forall k,$ and $0 < \sum_{k=1}^{N} \mu_{ik} < N, \forall i.$

The algorithm [16] proceeds as follows:

1) Pick the initial means m_i, i = 1, · · · , c. choose value for fuzzifier p and threshold ε. Set the iteration counter t = 1.
2) Repeat Steps 3-4, by incrementing t, until $|\mu_{ik}(t) - \mu_{ik}(t-1)| > \varepsilon$.
3) Compute μ_{ik} by eqn. (5) for c clusters and N data objects.
4) Update means m_i by eqn. (4).

2.4 Possibilistic c-Means (PCM) Clustering Algorithm

It partitions a set of N objects $\{x_k\}$ into c clusters by minimizing the objective function

$$J = \sum_{i=1}^{c}\sum_{k=1}^{N}(t_{ik})^q ||x_k - m_i||^2 + \sum_{i=1}^{c}\eta_i \sum_{k=1}^{N}(1 - t_{ik})^q \tag{6}$$

where $1 \leq q < 1$ is the fuzzifier, m_i is the ith cluster center, $t_{ik} \in [0, 1]$ is the typical membership of the k^{th} pattern, η_i are suitable positive integers and ||.|| is the distance norm, such that

$$m_i = \frac{\sum_{k=1}^{N}(\mu_{ik})^q x_k}{\sum_{k=1}^{N}(t_{ik})^q} \tag{7}$$

and

$$t_{ik} = \frac{1}{1 + \left(\frac{d_{ik}^2}{\eta_i}\right)^{\frac{1}{q-1}}} \tag{8}$$

and

$$\eta_i = K\frac{\sum_{k=1}^{N} t_{ik}^q d_{ik}^2}{\sum_{k=1}^{N} t_{ik}^q} \tag{9}$$

here typically K is chosen to be 1. The main steps of the PCM algorithm [17] are as follows:

1) Pick the initial means m_i, i = 1, · · · , c. choose value for fuzzifier p and threshold ε. Set the iteration counter it = 1.
2) Repeat Steps 3-4, by incrementing it, until $|t_{ik}(it) - t_{ik}(it-1)| > \varepsilon$.
3) Compute t_{ik} by eqn. (8) for c clusters and N data objects.
4) Update means m_i by eqn. (7).

2.5 Fuzzy-Possibilistic c-Means (FPCM) Clustering Algorithm

It partitions a set of N objects $\{x_k\}$ into c clusters by minimizing the objective function

$$J = \sum_{i=1}^{c} \sum_{k=1}^{N} \left(\mu_{ik}^{p} + t_{ik}^{q}\right) ||x_k - m_i||^2 \tag{10}$$

subject to the constraints p > 1, q > 1, $0 \leq \mu_{ik}, t_{ik} \leq 1$, and

$$\sum_{i=1}^{c} \mu_{ik} = 1, \forall k \tag{12}$$

and

$$\sum_{k=1}^{N} t_{ik} = 1, \forall i \tag{12}$$

and

$$m_i = \frac{\sum_{k=1}^{N} \left(\mu_{ik}^{p} + t_{ik}^{q}\right) x_k}{\sum_{k=1}^{N} \left(\mu_{ik}^{p} + t_{ik}^{q}\right)} \tag{13}$$

here μ_{ik} is the fuzzy membership value given in eqn. (5) and t_{ik} is the typical or possibilistic membership value given in eqn.(8), p and q are fuzzifiers.

The main steps of the FPCM algorithm [18] are as follows:

1) Pick the initial means m_i, i = 1, · · · , c. choose value for fuzzifier p, q and threshold ε. Set the iteration counter it = 1.
2) Repeat Steps 3-4, by incrementing it, until $|\mu_{ik}(it) + t_{ik}(it) - \mu_{ik}(it - 1) + t_{ik}(it - 1)| > \varepsilon$.
3) Compute μ_{ik} by eqn. (5) and t_{ik} by eqn. (8) for c clusters and N data objects.
4) Update means m_i by eqn. (13).

2.6 Imputation of Missing Values

Initially, all missing values in T are replaced by the estimation given by row (gene) averages to obtain a complete matrix. Specially, this step of gene average substitution, performed in all clustering methods, provides the possibility of contributing the maximum number of genes for estimating the missing values. Then any one of the above mentioned clustering algorithms are executed on this complete matrix. The missing values are imputed by making use of the weighted mean of the values of the corresponding attribute over all clusters. The weighting factors are the membership degrees u_{ik} of a gene x_k to the i^{th} cluster. The missing gene expression value x_{kj} is imputed by:

$$x_{kj} = \frac{\left(\sum_{i=1}^{c} u_{ik}^{l} v_{ij}\right)}{\sum_{i=1}^{c} u_{ik}^{l}} \tag{14}$$

where u_{ik} is the membership value of k^{th} gene in the i^{th} cluster. v_{ij} represents value of j^{th} attribute of mean of i^{th} cluster and l is the fuzzifier. For hard c-mean clustering membership values are either 0 or 1.

The main steps of the imputation algorithm is as follows:

1) Initially all missing values in T are replaced by the estimation given by row (gene) averages for obtaining a complete matrix.
2) Apply any one of the above mentioned clustering algorithm to cluster genes.
3) Estimate missing values by using eqn.(14) with the means obtained from clustering result.
4) Repeat steps 1 and 2 for different number of clusters.

3 Experimental Results

The above mentioned different partition-based clustering algorithms are compared with the previously developed KNNimpute and SKNNimpute methods by imputation of microarray data. Data sets used in this work are selected from publically available microarray data. Three microarray data sets are used: cluster analysis and display of genome-wide expression patterns (data 1) [19], Genomic expression programs in the response of yeast cells to environmental changes (data 2) [20] and the transcriptional program in the response of human fibroblast to serum (data 3)[21]. The metric used to assess the accuracy of estimation is Root Mean Squared (RMS) error which is calculated as follows:

$$RMS_{error} = \frac{\sum_{h=1}^{n} (R_h - I_h)^2}{n} \tag{15}$$

where R_h is the real value, I_h is the imputed value, and n is the number of missing values.

Before any further process, each data set is preprocessed for the evaluation, by removing rows and columns containing missing expression values greater than 50% and rest are replaced by row average values, yielding complete matrices. For every data set between 1 and 20% of the data are deleted at random to create test data set. Each method is then used to recover the introduced missing values for each data set, and the estimated values are compared to those in the original data set.

Every clustering method is executed for c = 5 to 50, where c is the number of clusters. The experiments show that for c > 50 the clustering results detoriates. The value of fuzzifier is varied from 1.2 to 2. For every clustering method best result (i.e. minimum RMS error) is taken for different values of fuzzifier as well as for different values of number of clusters (c).The result is shown for different rate of missing entries present in every data set.

The efficiency of different partition-based clustering methods mentioned here are compared with the KNNimpute and the SKNNimpute methods by applying them to three microarray data sets with different missing rates. Both KNNimpute and SKNNimpute methods require the value of k which is the number of nearest neighbors used in imputation. When k is between 5 and 20, they have given good performances. Accordingly, minimal RMS errors of these two methods are shown by

varying k between 5 to 20 in every data set with different rates of missing values. In Table 1, prediction accuracies of different clustering methods are shown for different number of clusters. In Table 2, only best results are shown for data 2 and Data 3 using different clustering algorithms.

In figure 1, it is found that c-mean has given best results compared to all other partition-based clustering algorithms mentioned here and also with respect to KNNimpute and SKNNimpute methods for all different rates of missing entries in data 1 and data 2. FCM, PCM, and FPCM clustering methods also have given better results with respect to KNNimpute and SKNNimpute methods for all cases in data 1 and data 2. For data 3, FCM gives best results for all rates of missing. The other clustering methods have also given better results compared to KNNimpute and SKNNimpute methods for data 3.

Table 1. Comparative Performance Analysis of Different Clustering Methods on Data 1

Rate of data missing (%)	No. of clusters	Prediction Accuracy			
		c-Mean	FCM	PCM	FPCM
1	5	0.484	0.523	0.519	0.523
	10	0.474	0.522	0.524	0.483
	20	0.459	0.502	0.518	0.448
	30	0.45	0.499	0.522	0.454
	50	0.437	0.563	0.523	0.443
5	5	0.534	0.531	0.533	0.543
	10	0.47	0.529	0.529	0.494
	20	0.46	0.527	0.527	0.463
	30	0.452	0.502	0.525	0.464
	50	0.441	0.503	0.525	0.503
10	5	0.544	0.555	0.543	0.531
	10	0.501	0.524	0.513	0.494
	20	0.474	0.523	0.513	0.473
	30	0.463	0.515	0.524	0.502
	50	0.441	0.564	0.535	0.532
20	5	0.583	0.576	0.587	0.594
	10	0.556	0.563	0.542	0.542
	20	0.494	0.547	0.542	0.502
	30	0.463	0.542	0.564	0.524
	50	0.468	0.548	0.564	0.524

Table 2. Best Performance of Different Clustering Methods on Data 2 and Data 3

Data Set	Rate of data missing(%)	Prediction Accuracy							
		No. of clusters	c-Mean	No. of clusters	FCM	No. of clusters	PCM	No. of clusters	FPCM
Data 2	1	50	0.51	30	0.59	10	0.7	20	0.61
	5	50	0.53	30	0.62	10	0.71	10	0.61
	10	50	0.54	25	0.64	20	0.73	10	0.62
	20	50	0.55	30	0.65	20	0.74	20	0.65
Data 3	1	30	0.74	15	0.68	50	0.8	10	0.73
	5	20	0.75	20	0.7	15	0.8	25	0.76
	10	30	0.76	10	0.71	10	0.85	30	0.75
	20	30	0.8	10	0.75	10	0.86	30	0.8

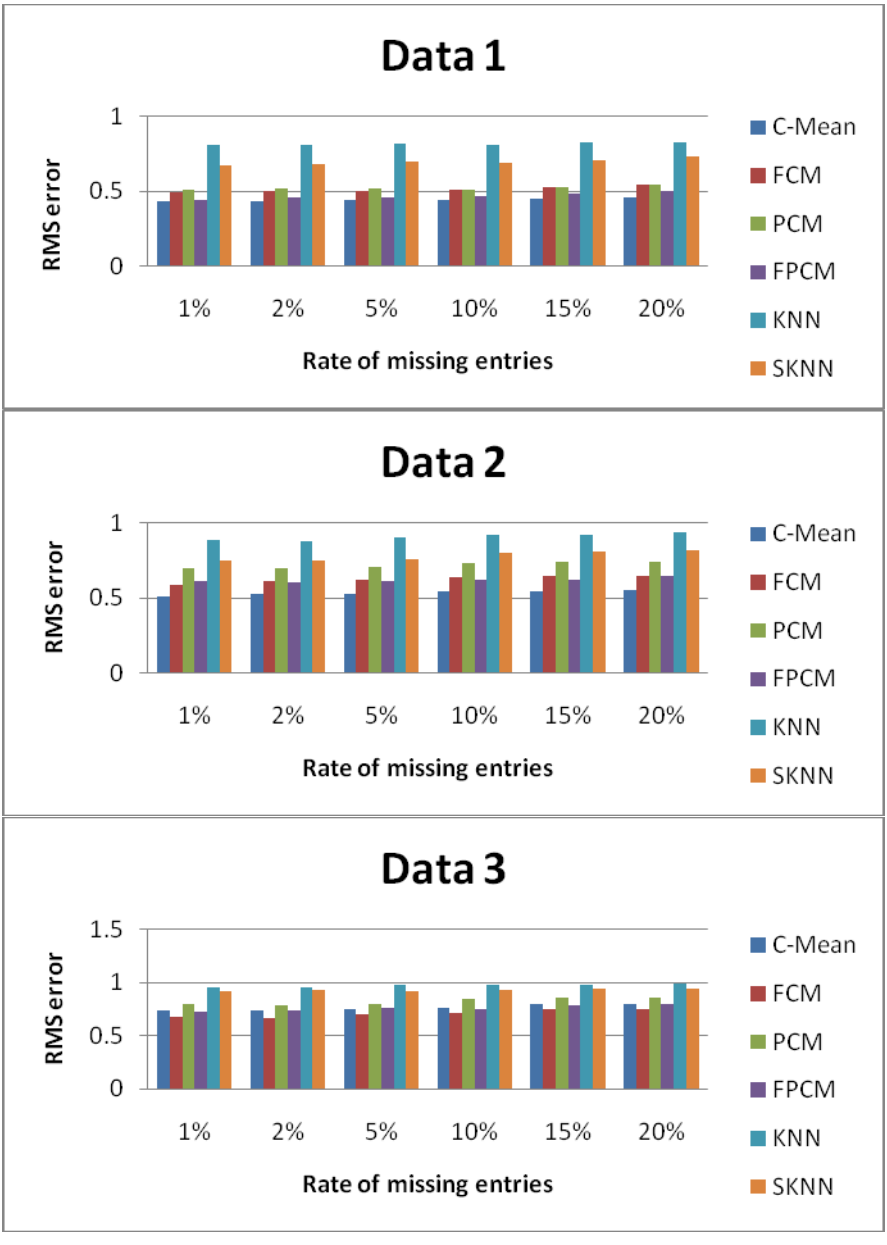

Fig. 1. Comparison of accuracy of Different clustering methods with KNNimpute and SKNNimpute methods for three types of data sets over 1 to 20% data missing. The accuracies are evaluated by RMS error.

4 Conclusion

In this paper, the performance accuracy of different partition-based clustering algorithms for missing value estimation in microarray data sets are compared with KNNimpute and SKNNimpute methods. The experimental results show that in all cases clustering methods have given better results than KNNimpute and SKNNimpute methods in terms of RMS error. So, it can be concluded that clustering methods are also very effective for missing value estimation in microarray gene expression data.

References

1. Schulze, A., Downward, J.: Navigating gene expression using microarrays - a technology review. Nat. Cell Biol. 3, E190–E195 (2001)
2. Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson, J.J., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Staudt, L.M.: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403, 503–511 (2000)
3. Raychaudhuri, S., Stuart, J.M., Altman, R.B.: Principal component analysis to summarize microarray experiments: application to sporulation time series. In: Pac. Symp. Biocomputing, pp. 455–466 (2000)
4. Alter, O., Brown, P.O., Bostein, D.: Singular value decomposition for genome-wide expression data processing and modeling. Proc. Natl Acad. Sci. USA 97, 10101–10106 (2000)
5. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Bostein, D., Altman, R.B.: Missing value estimation methods for DNA microarrays. Bioinformatics 17, 520–525 (2001)
6. Kim, K.Y., Kim, B.J., Yi, G.S.: Reuse of imputed data in microarray analysis increases imputation efficiency. BMC Bioinformatics 5(160) (2004)
7. Oba, S., Sato, M.A., Takemasa, I., Monden, M., Matsubara, K.I., Ishii, S.: A bayseian missing value estimation method for gene exression profile data. Bioinformatics 19, 2088–2096 (2003)
8. Wang, X., Li, A., Jiang, Z., Feng, H.: Missing value estimation for DNA microarray gene expression data by support vector regression imputation and orthogonal coding scheme. BMC Bioinformatics 7, 1–10 (2006)
9. Wong, D.S.V., Wong, F.K., Wood, G.R.: A multi-stage approach to clustering and imputation of gene expression profiles. Bioinformatics 23, 998–1005 (2007)
10. Friedland, S., Niknejad, A., Chihara, L.: A simultaneous reconstruction of missing data in DNA microarrays. Linear Algebra Appl. 416, 8–28 (2006)
11. Kim, H., Golub, G.H., Park, H.: Missing value estimation for DNA microarray gene expression data: local least squares imputation. Bioinformatics 21, 187–198 (2005)
12. Sehgal, M.S.B., et al.: Statistical neural networks and support vector machine for the classification of genetic mutations in ovarian cancer. In: IEEE CIBCB 2004, USA (2004)
13. Sehgal, M.S., et al.: K-ranked covarience based missing values estimation for microarray data classification. In: HIS (2004)

14. Au, W.-H., Chan, K.C.C., Wong, A.K.C., Wang, Y.: Attribute clustering for grouping, selection, and classification of gene expression data. IEEE Trans. on Computational Biology and Bioinformatics 2(2) (2005)
15. Tou, J.T., Gonzalez, R.C.: Pattern recognition principles. Addison-Wesley, London (1974)
16. Bezdek, J.C.: Pattern recognition with fuzzy objective function algorithms. Plenum Press, New York (1981)
17. Krishnapuram, R., Keller, J.: A possibilistic approach to clustering. IEEE Trans. Fuzzy Syst. 4(3), 393–396 (1993)
18. Pal, N.R., Pal, K., Bezdek, J.C.: A mixed c-means clustering model. In: IEEE Int. Conf. Fuzzy Systems, Spain, pp. 11–21 (1997)
19. Eisen, M., Spellman, P., Brown, P., Bostein, D.: Cluster analysis and display of genome wide expression patterns. Proc. Natl Acad. Sci., USA 95, 14863–14868 (1998)
20. Gasch, A., Spellman, P., Kao, C., Carmel-Harel, O., Eisen, M., Storz, G., Bostein, D., Brown, P.: Genomic expression programs in the response of yeast cells to environmental changes. Mol. Biol. Cell. 11, 4241–4257 (2000)
21. Iyer, V.R., Eisen, M.B., Ross, D.T., Schuler, G., Moore, T., Lee, J.C.F., Trent, J.M., Staudt, L.M., Hudson, J.J., Bogosk, M.S., et al.: The transcriptional program in the response of human fibroblast to serum. Science 283, 83–87 (1999)

MACREE – A Modern Approach for Classification and Recognition of Earthquakes and Explosions

K. Vijay Krishnan, S. Viginesh, and G. Vijayraghavan

Anna University, Chennai – 600025, Tamil Nadu, India
{vkrish.krishna,vigineshsankararaman}@gmail.com

Abstract. Though many systems are available for discrimination between earthquakes and explosions, our introduces new advances and some rudimental results of our ongoing research project. To discriminate between earthquakes and explosions, temporal and spectral features extracted from seismic waves, additional some seismological parameters (such as epicenter depth, location, magnitude) are crux for rapid and correct recognizing event sources (earthquakes or explosions). Seismological parameters are used as the first step to screen out obvious earthquake events. Fourier transforms (FFT), chirp-Z transforms, wavelet transforms have been conducted and some prominent features are acquired by present experimental dataset. In some experiments, wavelet features plus support vector classification (SVC) have reached very high correct recognition rate (>90%). This proposed paper can be used in evolving scenarios.

Keywords: Classification, Recognition, Earthquake, Explosion, Tempo-Spectral features, Support Vector Machines (SVC).

1 Introduction

The research on seismic signal processing, analysis, and further discrimination of earthquakes and explosions plays a fundamental role in the development of seismology, and is also indispensable for public welfare and world peace. Modern digital seismographs may record seismic waves of earthquakes and significant explosions occurring sequentially or simultaneously. These sequential or simultaneous occurring characteristics would be harmful to properly explain the recorded seismic waves and might beget some false conclusions. So it is very meaningful to separate earthquake events and explosion events which may occur sequentially or simultaneously from recorded seismic waves. The separation of simultaneously occurring earthquake and explosion events is one aim of our next researches, and should be investigated in further researches by some special signal processing techniques such as independent component analysis [1].

We are presently focused on the separation of sequentially occurring earthquake and explosion events or unrelated earthquake and explosion events. Sequentially occurring continuous or intermittent events have been processed, and saved in seismic wave files so that each file contains only one event, earthquake or explosion.

N. Meghanathan et al. (Eds.): Advances in Computing \& Inform. Technology, AISC 177, pp. 49–56.
springerlink.com

2 Discriminative Features

It has been shown [2][3] that the hypocenters of earthquakes are almost deeper than those of explosions. This may impact the travel ways of seismic waves. In addition, the origin mechanisms of the 2 event types differ essentially. Many different temporal and spectral features (the ratio of different magnitude scales [4], seismic phase, P-wave initial arrival time, the direction of P-wave initial arrival, the ratio of P/S-wave magnitude values, relationship of waves, complexity of wave, ratio of spectrums, cestrum, instantaneous spectrum, etc) have been proposed and investigated, but no ideal feature(s) is widely accepted due to the problem's complexity and proposed and some encouraging rudimentary results have been acquired [5]. Initially by Fourier transform (FFT), overall spectrum layout is acquired. Fourier transform transfers a time domain signal into frequency domain. In the discrete form, its definition is as following:

$$X(e^{j\omega}) = \sum_{n=0}^{N-1} x(n) e^{-j\omega nk} \quad -1, 0 \le k \le N$$

Where $x(n)$ is the discrete time signal or sampled continuous time signal with length N, $\omega = 2\pi / N$ is the angular frequency, $X(e^{j\omega})$ is the Fourier Spectrum. Chirp-Z transform is a classical algorithm with thinning spectrum [6]. This algorithm is formulated from discrete Fourier transform. For N - points length time signal $x(n)$, the Chirp-Z transform is defined as [7]:

$$X(z_r) = CZT[x(n)] = \sum_{n=0}^{N-1} x(n) z^{-n} = \sum_{n=0}^{N-1} x(n) A^{-n} W^{nr}$$

Where $A = A_0 e^{j}, W = W_0 e^{-j\varphi}$; A, W both is Real numbers. If $A0 > 1$, the integral or accumulative path of CZT is outside the unit circle; otherwise, the path is inside the unit circle. If $W_0 > 1$, the path rotates inwards; otherwise, rotates outwards. If $A_0 = W_0 = 1$, the path is an arc along the unit circle. Θ_0 is the onset angular frequency, φ is the sampling interval also known as angular increment.

Let $\theta_r = \theta_0 + r\varphi_0$, $r = 0, 1, \ldots, M-1$ are interested points of frequency range. If $\theta_0 = 0$, $M = N$, Chirp-z transform and discrete Fourier transform are the same. If $\varphi_0 = 2\pi / M < 2\pi / N$, *CZT* sampling spectrum X (z r) is finer than DFT, it may acquire more precise spectrum characters.

The result [5] of an experiment for discriminating 40 earthquakes and 40 explosions with 0.01 Hz-5 Hz frequency range are listed in Figure 1. In this figure, abscissa axis represents the logarithm value of average energy (normalized), the ordinate axis means the logarithm value of dominant frequencies; and **Red circles represent explosion events, Blue asterisks represent earthquake events.**

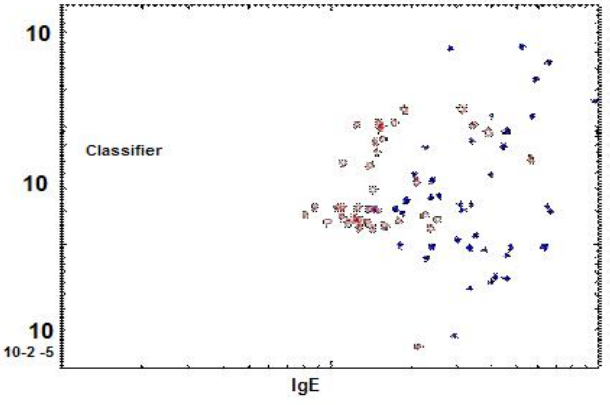

Fig. 1. Rudimentary result [5] of chirp-z transforms

This **means** that appropriate$_{n=0}$ window length of seismic signal may also be an important role in the classification of earthquake and explosion. But these seemingly encouraging results are very limited when applied to more other sites events.

Features extracted from chirp-z transform and wavelet transform are actually temporal-spectral quantities. How to combine theses temporal-spectral features with classical event features, wave temporal and spectral features for acquiring more robust classification and recognition result is imperative but challenging due to heterogeneity of earth structure and complexity of event behaviors. Thus, location comparability and magnitude scale comparability are fundamental requirements for high accurate explosion recognition.

3 Recognition Features

In pattern recognition field, recognition and classification are often interchangeably used though their meanings are subtle different from each other. Strictly speaking, recognition is identified a new sample as one of some several presumable classes, and classification is the process of designing some rule(s), then designating each of a group of samples to one of correspondent class. For a pattern recognition system, recognition is often considered in testing and practical application phases, and classification is often considered in training and learning phases. A typical pattern recognition system is sketched in Figure 2.

The transform of 4-wavelet packet has also already been investigated to extract 3 types of wave features - energy ratio, Shannon entropy and logarithmic energy entropy:

$$E_{wt}(i\ ) = 100 * \sum_{n=1}^{N} x^2(i,j) / \sum_{m=1}^{J} x^2(n,m)$$

$$E_{shannon}(i) = -\sum_{j=1}^{J} x^2(i,j)\log(x^2(i,j))$$

$$E_{\log}(i) = -\sum_{j=1}^{J} \log(x^2(i, j))$$

These features are supplied to a classifier of v-SVC (support vector classifier) for verifying the capabilities of these features. In three approaches of pattern recognition: syntactic, neural and statistical, the last one is main force in many practical recognition applications. In the statistical approach of pattern recognition, the main focus of researches is acquiring some technique for best generalization of decision rules which derived from training samples in experimental data sets. This approach requires a powerful computational capability, demanding some flexible use of numerical programs for studying the data set as well as tools for evaluating the data analysis procedures themselves. As many new techniques are still being proposed in the literature, an easy, robust one.

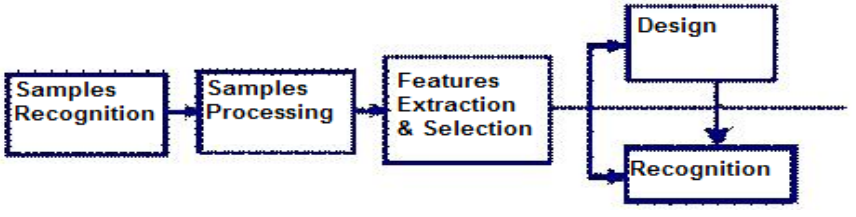

Fig. 2. Sketch of Pattern Recognition system

The results [8] showed that the feature of Shannon entropy is the best candidature to discriminate earthquake and explosion among the above three features. Classifications by v -SVC are carried out for more elaborate recognition tests. The results show that window length is also an important factor for recognition rate. The recognition rates of several different window lengths are ranged from 81% to 98%. The best window length is 2000 sampling points which achieves 98% programming tool or platform is needed that enables a fast and flexible algorithm implementation. Hereby the use of a widely available numerical toolset like Matlab may be profitable for both, the use of existing techniques, as well as for the study of new algorithms. Moreover, because of its general nature in comparison with more specialized statistical environments, it offers an easy integration with the preprocessing of data of any nature. This may certainly be facilitated by the large set of toolboxes available in Matlab. So the recognition algorithms are current implemented in Matlab.

Because of abounding pattern recognition algorithm existing in literature, each different algorithm may be best for its suitable situation and sample data structure. But none of any single algorithm is good for any structural samples. So, basing upon the careful analysis of present problem's sample structure, several typical algorithms have been selected to accommodate the classification of earthquake and explosion. They are Fisher's classifier, a classic linear classifier which suitable to linear classifiable problem; ICHAM (Improved Continuous Hamming's Method) classifier, which suitable to interleaving conglobation sample structure problem; general linear classifier, different from Fisher's by no need to calculate covariance and ever applicable to some

non-linear classifiable problem if incurring errors is acceptable; Parzen's classifier, which not needs any presumption and completely learned from data; and a v-SVC(support vector classifier), which entirely is in a new frame and theoretically can be used to formulate almost any shape of delimitating boundary line for any structure's samples, linear classifiable and non-linear classifiable.

The parameters [8] for v-SVC are set as follows:

ν =0.5; Kernel is Sigmoid

$$k(x_i, x_j) = \tanh((\gamma x_i - x_j + c))$$

Where x_i, x_j are $i-thj-th$ sample vector(s) respectively. Parameters in Sigmoid kernel: $\gamma = 1/100$, $c = 0$.

4 Decision Support System

Construction of suitable decision support system is the best solution for explosion recognition and earthquake vs. explosion events classification. The schematic layout for the decision support system of recognizing natural earthquake events and explosion events is displayed in Figure 3 and a snapshot of user operation interface is displayed in Figure 4. The developing platform of this system is MS Visual Studio 2008, programming language is C#. Most seismic signal processing algorithms are coding and developing in Matlab (v7.1) and packed as dll (dynamic-link library) files.

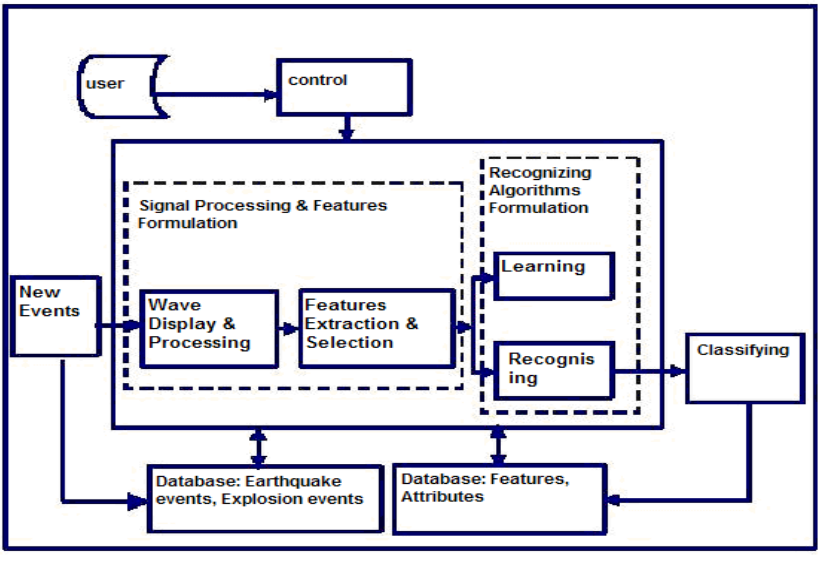

Fig. 3. Layout of the decision support system

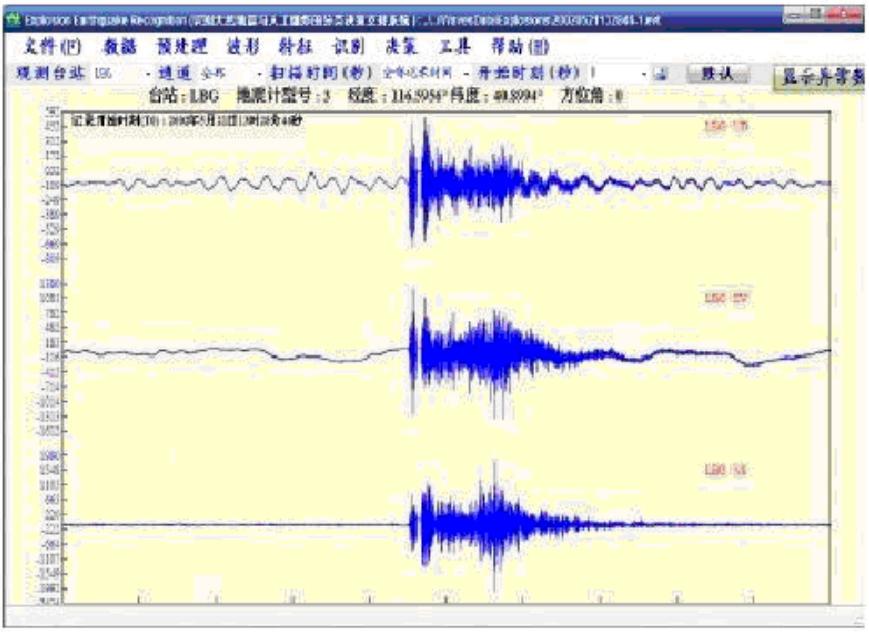

Fig. 4. Interface of the decision support system

E Event wave data that stored in popular seismic data formats such as evt, sed, sac, and txt can be read in and processing in current developing phase, more data formats, especially seed (Standard for the Exchange of Earthquake Data), are planning to add into the system. Prevailing seismic wave data are stored as one file for one event, which may contain more than 100 observatory stations, and generally each station has 3 channels (UD, EW, NS). Thus, the size of one wave data file may larger than 500MB for some events.

Raw seismic wave records may contain a great deal of silent void data, outburst disturbing data, environmental noising data, device trending excursion, etc. So some appropriate data preprocessing must be applied before any meaningful signal processing and wave feature extraction.

Event features such as event magnitude(s) (Ms, Mb, etc), hypocenter depth, epicenter longitude/latitude, must be inputted into the system by user(s), and stored in an event catalog database. However, the raw event wave file cannot be directly stored in database due to its huge size (may > 500MB/single file). But the name of raw event wave file must be associated with event catalog, and can be simultaneously stored in the database.

Temporal features extracted from event waves shall be acquired by user's interacting operation with the waves in the software interface. P-wave initial arrive time, S -wave initial arrive time, P-wave initial magnitude, P-wave maximal magnitude, S-wave maximal magnitude etc, these temporal-related wave signal values can be observed and measured in seismic wave graph. Temporal features can be gotten or calculated from these measurements.

Spectral features are computation-consuming quantities. In order to utilize Matlab's powerful scientific computing capabilities, all spectral features extraction algorithms(chirp- Z, wavelet, Hilbert-Huang etc) are coded and completed in Matlab and these algorithms are packed in several DLL files for convenient called by the system's user.

Explosion recognition and classification of earthquake and explosion events are the core parts of the system. Several pattern classification algorithms (SVC, Fisher's, ICHAM, Linear, and Parzen's Classifier) have been implemented in the system. Of these algorithms, SVC (support vector classifier) is the most robust one for our present experiments.

5 The Working of Next Step

The main and core parts of system's software have basically been completed, but there are still many problems waiting to be solved. Algorithms' bugs may unceasingly be found, and algorithms' limitations also perhaps will be faced when more event data arriving and more experiments being conducted. So algorithms improvements and ever adaptations are always needed. Statistical decision module is also needed due to different pattern recognition algorithm may derive inconsistent conclusions for just one event source (explosion or earthquake).

For clarification, the workings of next step are briefly listed as follows:

- Refining user interface, wave operation, wave signal filtering;
- Debugging temporal, spectral, tempo-spectral feature extraction algorithms;
- Implementing features browse and update;
- Refining explosion recognition algorithm;
- Completing statistical decision and inference module.

References

[1] Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Wiley and Sons (2001)

[2] Tjøstheim, D.: Improved seismic discrimination using pattern recognition. Physics of the Earth and Planetary Interiors, 85–108 (1978)

[3] Bian, Y.J.: Application of genetic BP network to discriminating earthquakes and explosions. Acta Seismologica Sinica 15(5), 540–549 (2002)

[4] Bian, Y.J.: Application of Fisher method to discriminating earthquakes and explosions using criterion mb/Ms. Acta Seismologica Sinica 18(4), 441–450 (2005)

[5] Huang, H.M., Bian, Y.J., Li, R., Lu, S.J.: Discrimination of earthquakes and explosions using chirp-z transform spectrum features. In: Proceedings of the WRI World Congress on Computer Science and Information Engineering, vol. 7, pp. 210–214 (2009)

[6] Rabiner, L.R., Schafer, R.W., Rader, C.M.: The chirp z-transform algorithm. IEEE Transactions on Audio and Electro-acoustics (June 1969)

[7] Hu, G.S.: Digital Signal Processing Theory, Algorithms and Implementations. Qsinghua Unversity Press, Beijing (2003)

[8] Huang, H.M., Bian, Y.J., Lu, S.J., Jiang, Z.F., Li, R.: A research on seismic signal wavelet features of earthquake and explosion. Acta Seismologica Sinica (2010) (in press)
[9] Huang, H.M., Bian, Y.J., Lu, S.J., Li, R., Jiang, Z.F.: *v* SVC algorithm applied in earthquake and explosion recognition and the choice of window length. Seismological and Geomagnetic Observation and Research (2010) (in press)

Building Concept System from the Perspective of Development of the Concept

Cheng Xian-yi, Shen Xue-hua, and Shi Quan

School of Computer Science and Technology, Nantong University,
Nantong Jiangsu, 226019, China

Abstract. The concept system with rich content is the key to improve the performance of knowledge-based artificial intelligence knowledge system. And a sufficient number of concepts, rich in semantic association, to meet the multi-tasking and developed concept system are one of the major challenges of knowledge engineering. It is the fundamental goal of conceptualization of knowledge, too. In this paper, for the study of natural language processing, from the perspective of development of the concept, a framework is proposed to building concept system.

Keywords: Natural Language Processing, concept system, ontology, semantics, knowledge.

1 Activities of Senior Intelligence

The Original intention of artificial intelligence is to simulate person's intelligence. However, because we know little about the generating process of living beings' (especially human) intelligent behavior, we can only explore the problem that is easier to solve by existing tools. The knowledge expressions that are used by existing systems of artificial intelligence are all aimed at particular application. That is problem oriented. Both Symbolisms and Connectivism are all approximation of knowledge representation. With the development of artificial intelligence in deep and application, many problems put forward a challenge for the existing technology of artificial intelligence. The applications that had been paid more attention are as the following several aspects [1].

(1)Although we have find much new applicable technology about machine translation, automatic abstract, question-answering system and so on in the field of natural language processing, it is the calculation of concept which is not the keyword matching even in a small category, such as what is the meaning of "Li go tomorrow".

(2)In the field of problem solving, being different from weak method and the method of expert system which put the state space search as its core ,many new applications put forward higher request to the method which is based on knowledge ,and it outstands more and more knowledge intensive features. For example, the low temperature disaster happened in the middle and low areas of the Yangzi River before the Festival Spring in 2008 years. Its rescue involved the weather, traffic, electric power, medical

N. Meghanathan et al. (Eds.): Advances in Computing \& Inform. Technology, AISC 177, pp. 57–65.
springerlink.com

treatment, civil affairs, and many other fields of knowledge. It no longer meets the reasoning realization of the classic expert system but the concept association.

(3)In the field of Web information processing, for the extraction of the interested information and the filter of the harmful information online, we have developed the text mining, public opinion monitor, vertical search and other new demand, and the user hope to put the Web as a problem solver but not a resource pool.

(4)In the filed of scene analysis, for the computer vision and pattern recognition, we apply ourselves to the understanding of the image consistently and we hope to find the public emergencies, to recognize and alarm the internal theft of self-service bank, to find the old man's abnormal step appearance by accompanying robot and to forecast the production of food, and so on, through the image analysis which is captured by camera or remote sensing satellite.

(5)In the field of distributed computing, such as MAS (Multi-Agent Systems) and the robot football, it is not simple the task distribution or data distribution but needs fusion, collaboration and coordination.

(6)The challenge of NP problem. For example, the problem of protein folding is the process to study how a string of amino acids could fold almost in a moment and form a kind of very complex three-dimensional structure protein. It would take 10^{127} years for a super computer which used the verisimilitude for this project calculation to look for the final folding form of a short sequence which consists of only 100 amino acids [2]. There are a lot of problems like this.

The common between previous six problems and the problems of traditional artificial intelligence is the needing of knowledge, but they have the absent distinction. To solve them, senior intelligence system has to deal with a lot of variables. It is not a stack of simple concept (knowledge), but needs to understand the dependent of concept and highlights the knowledge hidden behind them.

In order to realize the senior intelligence, we need to construct a knowledge system which is rich in content, used neatly, explicit and can be stated. It is concept system that supports the system.

2 The Development of Concept System

In daily life, a seemingly simple common sense issue may involve a lot of complex association of concept, and these concepts and the connections between the concept is gained in the process of people's cognitive development. In the process of cognitive development, a child gets new concepts constantly and finds the dependent relationship between them and the existing concepts. The child's concept structure become complicated and complete gradually through such a long process of accumulation.

Here is a list of several influential concept systems:

2.1 FramNet

FramNet treats the frame as its core and it is based on the real corpus. It puts many lemmas that have the same semantic roles in the same frame and uses frame elements

which have individual character to describe the protean semanteme of nature language. Then it reveals the various semanteme of each word in each meaning and the possibility of syntactic integrating by the marked sentence[3]. For example, the word "hit" can be expressed as "zonk" and can also be treated as "create an accident or bad effect". How can we distinguish its meanings? Generally speaking, a word's different meanings are associated with the different meaning framework that the word participates in. When the meaning of a word is based on a particular frame, we would say that the word has activated a framework. Accordingly, the word "hit" activates a "hit the target" framework in a certain context environment and it also may activates a specific "cause harm" framework in another context environment[2].

2.2 Semantic Web

In brief, semantic Web is a kind of intelligent network that can understand human language. It can not only understand the human language, but also can make the communication between people and computer as easy as the communication between people. The core of it: it can add the semanteme (Meta data) that the computer can understand by the documents on the World Wide Web (such as the HTML), so that the entire Internet could become a general medium of information exchange. We can improve the resources' usability and usefulness of the World Wide Web and its interconnection by the following methods.

Although semantic web is a better network, its realization is a complex and huge engineering. The realization of semantic web is finished by XML Language and framework of resource description (RDF). XML is a tool used to define markup language. It includes XML declaration, DTD (document type definition) which is used to define the grammar of language, the detail instruction which is used to describe marks and the document itself. The document itself contains marks and content too. RDF is used to express the content of webpage.

The system structure of semantic web is shown in figure 1.

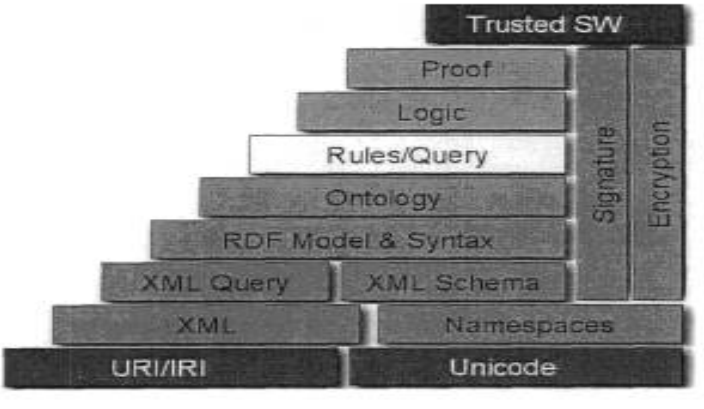

Fig. 1. The system structure of semantic web

The first layer: Unicode +URI (international code+ uniform resource logo). It is the foundation of the whole system structure. Unicode is a character set and is the code that is responsible for handling resources. URI is responsible for the identification of resources .it makes the precise retrieval of information and intelligence possible.

The second layer: xML + NS (Namespaee) + XMLSehema (extensible markup language + name domain + program of extensible markup language). It is responsible for representing the content and structure of data from grammar. XML uses a set of elements which is programmed prior to mark on the data and provides convenience for the computer processing. Name domain can distinguish the data elements' belong and can transform the synonyms between different name domains.

The third layer: RDF + RDFSchema (framework of resource description+ framework program of resource description). It is responsible for providing semantic model to describe the content and structure of information on the Web. RDF is a language of describing the information resources on the Web and its goal is to establish a framework which is used for the coexistence of a variety of metadata standard. The framework can make full use of all sorts of metadata and start exchange and use of data which is based on Web. RDFSchema uses the expression system that machine can understand to define the vocabulary of description resources.

The forth layer: Ontologyvoeabula (the collection of ontology vocabulary). It is responsible for the definition of sharing knowledge and the description of the relationship between the various resources.

The fifth floor: Logic. It is responsible for providing the inference rule of justice and logic and the basis for the intelligence service.

The sixth layer: Proof. It provides support for the signature of mutual validation and data exchange among intelligence agents.

The seventh layer: Trust. It provides trust guarantee.

The reasoning from the fifth layer to the seventh layer is on the based of the nether 4 layers.

The system structure of semantic web is under construction and the research of this system structure has not formed a logical description and theory system which is satisfying and strict at current international scope.

2.3 Concept Map

Concept map is a more modern method of knowledge representation. It was first proposed by John f. Sowa in 1984 and was a model of expressing language in network [6]. Concept map is a directed and connected graph that is represented by graph. It includes two nodes: concept node and concept relation node. The direction of the arc represents the relationship between concept knot and concept relation knot. Concept node represents a concrete or abstract entity in the field of question .concept relations note points out a kind of relationship that involves in one or more concept nodes.

In the concept map, concept node is represented with a rectangular and concept relation node is represented with an ellipse. Directed arc marks up the adjacent concept nodes of the concept relation node. Each concept map can represent a proposition and a typical knowledge base contains a lot of figures like this. For example, the concept map in picture 3 represents the proposition of "zhangsan gives lisi a red book ".

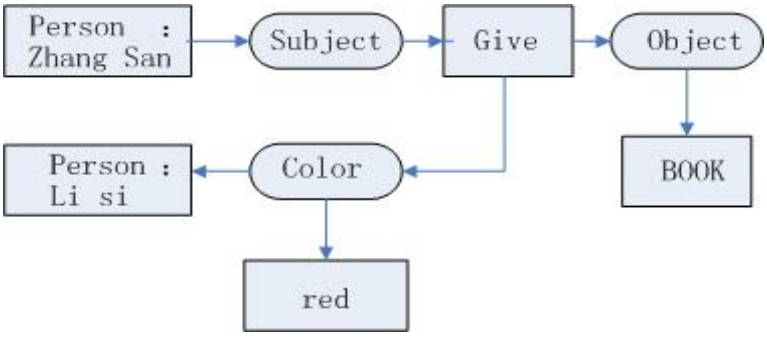

Fig. 2. Concept map examples

We can operate on concept map and create a new concept map. The formal rules of operation: copy, restrict, connect and simplify.

Concept map is a logic system which is based on semantic network. It is not only straightforward but also easy to operate for concept map to express the knowledge. It can produce new concept relevance and inference rule with the operating of concept map. In addition, the concept map can also create a mapping relationship with natural language directly. These advantages that concept map involves make it more suitable for expressing concept structure. Concept map has been favored by many researchers since it was putted forward and they have applied it in many different fields (such as knowledge engineering, information retrieval). Comparing with the classic method of knowledge representation, Concept map is more coincident with human's thinking and language habits. But it can only express some simple concept relations and is not applicable to express commonsense knowledge that contains complex concept structure. Conducting knowledge representing with concept map needs to analyze the structure of knowledge said, so its acquisition needs the participation of experts and it can not be gained automatically by intelligent system. In addition, for the solving of a complex question, reasoning based on concept map is easy to produce redundancy or cause the disaccord of reasoning result. Therefore, an intelligent system based on concept map can only do some simple problem solving and it is not competent for commonsense problem solving that contains a large amount of complex concept relevance.

2.4 Concept Lattice

Concept lattice is a complete concept hierarchical structure that reflects the contact between objects and attributes and the relations between generalization and specialization. Each node of Concept lattice is a formal concept and it consists of two parts: the denotation and the connotation. Denotation is the examples that concept covers and connotation is the description of concept which covers the common features of examples. In addition, concept lattice manifests the relationship between generalization and specialization of these concepts concisely and vividly by Hasse diagram. Concept lattice is regarded as a powerful tool for data analysis. The process of creating concept lattice from data set (which is called formal background in concept

lattice) is a kind of process of concept clustering in essence. However, concept lattice can be used for many tasks of machine learning. Concept lattice has been applied extensively and successfully since it is putted forward. For example, we can extract multiple types of knowledge from concept lattice such as some properties and rules [8]. We can organize and manage a lot of information effectively by the concept lattice in information retrieval.

We can represent knowledge by concept lattice. We can also add the new obtainment of concept to the existing concept lattice by the structure algorithm of concept lattice, so as to we can update knowledge base constantly. Concept lattice can describe the category and hierarchical relationships of concept lattice clearly. It also can extract many common characteristics or rules of practical examples and organize them effectively. However, concept lattice is not suitable for expressing concept which has dynamic characteristics. We can obtain common properties of relevant concepts by concept lattice that is the connotation of concept. It is not good at expressing knowledge of common sense because the concept relations of commonsense knowledge contains are too complex and the concept relations that concept lattice can describe are too simple—they are only some relations of generalization and specialization. In addition, concept lattice has some problems itself. Such as the constructing efficiency of concept lattice, if the constructing algorithm is inappropriate, the number of concept lattice nodes may grow according to index level and this will cause a large number of redundant data.

2.5 HowNet

HowNet is a common sense knowledge which regards the concept that a Chinese word or an English word represents as its object and treats the relations between concepts or attributes of concept as its basic content [4]. With the investigation and analysis about 6000 Chinese characters, HowNet abstracts more than one thousand mete-senses. Mete-sense is the minimum unit which is the most basic and whose meaning can't be segmented in HowNet. It is the basic factor to explain the dictionary of knowledge and other vocabulary entries are all defined by it. Computerization is the important characteristic of know nets. HowNet is oriented to computer and it is established by computer. It may be an intelligent component of computer in the future. As a knowledge system, HowNet is a net but not a tree in fact [5].

HowNet also focus on reflecting the relationships between concepts or various attributes of concept. HowNet teaches the knowledge network system of figure 3 to the computer clearly and makes knowledge operable for the computer.

In general, the relationships that HowNet described between mete-senses mainly include hyponymy and, synonymy, amtonymy, the part-whole relation, the attributes-host relation, the material–product relation, the agency/experience/ subject of relation-event relation, the recipient/content/subjacency etc event relation, the tool-event relation, the place–event relation, the time-event relation, the value-attribute relation, the entity-value relation, the role-event relation and the correlative relationship and so on.

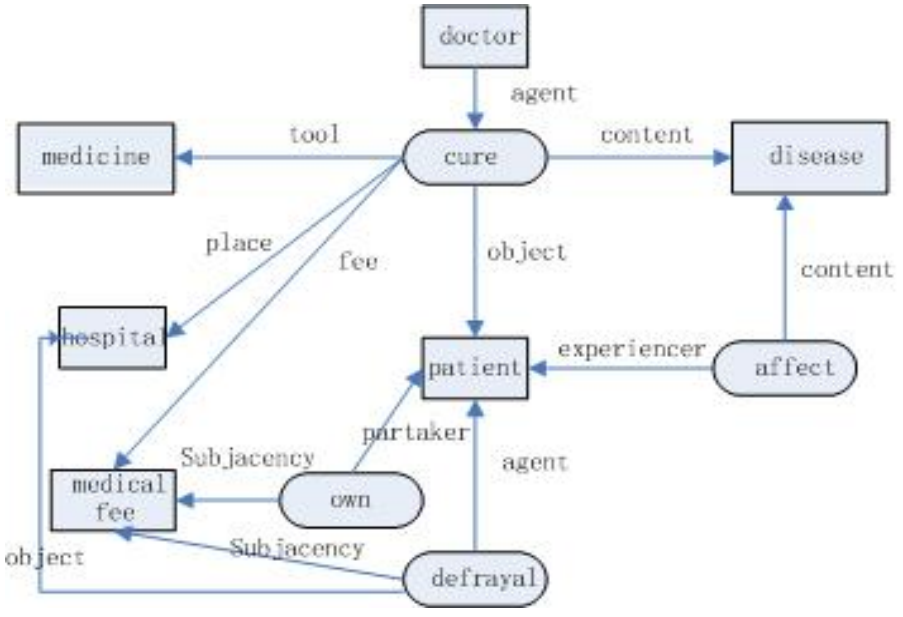

Fig. 3. Knowledge representations which is based on HowNet

2.6 HNC

HNC theory is a theory system about the understanding of natural language. It is a theory system that is based on semantic expression and mixes semanteme, grammar and pragmatic together. [7] HNC theory's goal is taking the association venation of concept as the main line, establishing a kind of natural language expression pattern and computer pattern of understanding and processing that can simulate the perception process of brain language and making the computer obtain the ability of digesting fuzzy.

HNC is based on the following two hypotheses:

Hypothesis 1: symbols of language space are different in thousands ways. However, there is only a kind of symbol of language concept space.

Hypothesis 2: role is in the internal and mutual space of all things and it must produce some role. It must be accompanying with a process or transfer before achieving final effect and it must be appearing a new relation or state after achieving final effect. New effect can induce new role and it recycles and recycles. This is the basic rule that all things exists and develops and this is also the basic rule of language expressing and concept reasoning (it becomes role and effect chain rule).

It introduces the language concept space. The meaning of the words can be mapped to the system of concept sign and it can be expressed by the combination of the concepts. HNC generalizes the symbolic expression of the natural language concept as the followings:

{Category string} {Hierarchical string} {Symbol of combination structure}

It means: the lambda expression of concept is consisted of category symbols, hierarchical symbols and combination symbols. Category string and hierarchical string constitute a lambda expression of concept primitive. Two or more concept primitives constitute a new concept by the combination of the combination structure symbol. The simple concept is consisted of a single concept primitive and the Compound concept is consisted of several concept primitives.

Example: Sun Zhongshan led the Xinhai revolution and overthrew the autocratic monarchy system which rules people thousands of years in china. It is great significant to the social progress, but it also fails to change the tragic fate of Chinese and society property of semi-colonial and semi-feudal in china.

Semantic tagging of the example's language concept space:

SG=R011X*20J#R0B1=<R411J>#R0B2=<!24R411X*21J[HE|]>+ReCS0jD1*20J#ReC= {PS*10J} +(lby)XY0*22J [7]

2.7 Ontology

Ontology is the clear specifications of conceptualization. It mainly includes four aspects:

(l)Conceptualization: abstract model of the objective world phenomenon.

(2)Clear: concepts and the relations among them are defined precisely.

(3)Formalization:accurate and mathematical description.It can be read by computer.

(4)Sharing: the knowledge that ontology reflects is recognized by its users.

Although there are many different ways about definition, different researchers have the unified understanding to ontology. They all take ontology as a semantic foundation of different subjects' (person, machine, software system, etc) communicating (dialogue, interoperability, sharing and so on) in the interior of field. That is ontology provides a consensus of clear definition. The goal of this consensus is mainly to service for the machine.

Many tools of ontology construction have been made in the past 10 years, such as from Ontolingua, OntoSaurus, WebOnto to Protege-2000, WebODE, OilEd, OntoEdit, KAON, Text-To-Onto, etc. These tools provide a graphical interface which is friendly and a mechanism of consistency check. With these tools, the users can concentrate on the organization of the ontology content without knowing the details of ontology description language and they also avoid a lot of mistakes and make it convenient for ontology construction. But, these tools only provide the editing function of ontology and what supports it is still the way of constructing ontology by hand. Even with these ontology editing tools, users still need to enter and edit the name, constraint, attributes, etc of each concept. The structure of the ontology is a five tuple ---O: = {C, R, Ao, F, S}. Here the C and R are two disjoint sets. The elements of C are called concepts and the elements of R are called relationship. Ao represents ontology axiom, F represents function and S represents example. We can see that the relation of ontology, example, function and axiom are all based on the basis of the

concept and the concept system plays an important role on the construction ontology from the structure of the ontology.

3 Conclusions

Concept system reflects the dependence of the concepts, which is the core of knowledge system, and has played a role lurking in the background. People's cognitive ability all comes from that. Thus the research of concept structure is helpful for improving the problem solving ability of knowledge system, especially the open field. The research of concept system is asked to solve three problems: how to represent and store concept system, how to construct concept system gradually and how to realize all kinds of intelligent task using concept system.

Acknowledgments. This dissertation is supported by National Natural Science Foundation of China (Grant No.: 61171132).

References

1. Wei, H.: Formal concept system of artificial intelligence. Science Press, Beijing (2011)
2. Li, Q.: FrameNet–a thesaurus engineering based on frame semantics. Science and Technology BBS 16, 39 (2005)
3. Jia, J.-Z., Dong, G.: The compare among Frame, WordNet, VerbNet. Library and Information Technology 11, 1682–1686 (2007)
4. Liu, Q., Li, S.-J.: The calculation of lexical semantic similarity based on HowNet. Computer Linguistics and Chinese Information Processing 7, 59–76 (2002)
5. Dong, Z.-D., Dong, Q.: HowNet (EB/OL) (September 23, 1999/March 06, 2004), http://www.keenage.com
6. He, W., Wei, H.: Study summarization of Concept structure. Computer Application and Software 27(1), 157–159 (2010)
7. Jin, Y.-H.: HNC Language Understanding Technology and Its Applications. Science Press, Beijing (2006)
8. Wang, D.X., Hu, X.G., et al.: Association rules mining on concept lattice using domain knowledge. In: The Fourth International Conference on Machine Learning and Cybernetics, pp. 2151–2154 (2005)
9. Wooldridge: (Shi, C.-Y., et al. (trans.)) The guidance based on many Agent systems. Publishing house of electronics industry, Beijing (2003)

E-Government Adoption: A Conceptual Demarcation

Rahmath Safeena and Abdullah Kammani

College of Computers and Information Technology
Taif University, Saudi Arabia
safi.abdu@gmail.com, akamani@acm.org

Abstract. The Information and Communication Technologies (ICT) are being increasingly used by various governments to deliver their services at the locations convenient to its citizens. E-government is a kind of governmental administration which is based on ICT Services. The essence of e-government is using information technology to break the boundary of administrative organizations, and build up a virtual electronic government. E-government initiatives are common in most countries as they promise a transparent, citizen-centric government and reduce operational cost. Emerging with E-government, theories and practices of public administration have stepped into a new knowledge era. E-government presents a tremendous impetus to move forward with higher quality, cost-effective, government services and a better relationship between citizens and government. This paper discusses the different issues, challenges, adoption factors for e-government implementation and, presents a conceptual demarcation on these factors.

Keywords: e-government, definitions, services, adoption factors, and challenges.

1 Introduction

The current society had a phenomenal transformation due to the advance of Internet. It has opened a new medium of communication for individuals, business, and government organization, providing more opportunities to communicate and get information in an entirely new way. It has made governmental information and services accessible in ways that could not have been conceived two decades ago [1]. In the past, government organizations paid little attention to service quality or responsiveness to clients, but this changed with the approach of E-Government. E-government refers to the use by state authorities of ICT, in particular, the Internet and web-based technology, to deliver information and services and to encourage civic participation [2]. E-government is simply a facility using Information Technology (IT) to deliver public services directly to the customer, where the customers are citizens, business or other government entity [3, 4]. This phenomenon of e-government is increasingly attracting the attention of community citizens including politicians, economists, decision and policy makers amongst others. It has improved managerial effectiveness, and promoted democratic values of public services. It has the promise of increasing

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 67–76.
springerlink.com

accessibility to information, enhancing efficiency and facilitation of greater access to government officials [5, 6]. It is the medium of delivering improved services to citizens, businesses, and other constituents of society through drastically changing the way governments manage information. However, the e-Government challenge is not a technological one. Rather, the challenge is to use technologies to improve the capacities of government institutions, while improving the quality of life of citizens by redefining the relationship between citizens and their government [1]. The development of e-government also means increased electronic co-operation within and among public organizations which even puts demands on development that is not technology oriented. The development towards e-government involves social changes of work roles, attitudes and new competence needs [7].

2 E-Government

E-Government initially began as an intra-governmental communication tool. Initially the government organizations developed websites with information, then developed to online transactions - which made the citizens to engage in online participation that connect citizens and decision-makers [8–11]. E-government represents a fundamental change in the whole public sector structure, values, culture and the ways of conducting business by utilizing the potential of ICT as a tool in the government agency [12]. The Internet is indeed the most powerful and popular means of delivering the services to the customers or citizens. Hence, Web sites have been employed as a platform for delivering a wide range of government services electronically. E-government websites help citizens to gain information on government processes and services and hence participating in democratic processes from anywhere at any time. E-Government improves the efficiency and effectiveness all government operations, with citizens, as well as with other organizations. E-government applications include online payment of tax, bills, filling and submission of applications for several purposes; e-voting etc. e-Government gives citizens more control on interaction with the government; citizens can avail of the governmental services from anywhere and anytime [1]. E-Government is considered as tool for easy administration of governmental activities. Its success depends on its vast usage and management of its infrastructure. Utilization of e-government will provide benefits to the management philosophy of governments. Thus the citizens can collaboratively participate in decision making [13, 14]. Initially E-Government incurs a great cost in building infrastructure but gradually it implementation results in vast savings towards government's activities. It also increases transparency, and reduce corrupt activities in public service delivery. Table 1 depicts E-Government defined by various related study in the near past.

Table 1. E-government Definitions

Citations	Definition
[16, 17]	E-government is defined as the use of ICT to make government more accessible, effective, and accountable.
[13, 18, 19]	E-Government refers to the delivery of [government] information and services online through the Internet or other digital means.
[20]	E-government refers to strategies, organizational forms and processes, as well as information technology employed so as to enhance access to and delivery of government information and services to citizens, businesses, government employees and other agencies.
[21, 22]	E-Government is the use of ICTs in public administrations combined with organizational change and new skills in order to improve public services and democratic processes and strengthen support to public policies.
[23]	E-government is the process of offering better government service to the public.
[13, 14]	E-government is defined as the combination e-administration and e-democracy to achieve the objective of balanced e-government.
[13, 18]	E-Government is the delivery of fast services to citizens, businesses, and other members of the society.
[24–26]	E-Government refers to the strategic application of ICT to "provide citizens and organizations with more convenient access to government information and services; and to provide delivery of public services to citizens, business partners and suppliers, and those working in the public sector"
[15, 27]	E-government is the continuous optimization of service delivery channel, citizen's participation and governance.
[15]	E-government can be defined as a way for governments to use the most innovative information and communication technologies, particularly web-based Internet applications, to provide citizens and businesses with more convenient access to government information and services, to improve the quality of the services and to provide greater opportunities to participate in democratic institutions and processes.

3 E-Government Services

Like any other electronic services, e-government also constitutes various types of services. According to Fang [15] different types of e-government services are categorized in to eight types. 1) Government-to-Citizen (G2C); 2) Citizen-to-Government (C2G); 3) Government-to-Business (G2B); 4) Business -to-Government (B2G); 5) Government-to-Employee (G2E); 6) Government-to-Government (G2G); 7) Government-to-Nonprofit (G2N); 8) Nonprofit-to-Government (N2G). Table 2 gives definition for these of e-government services.

Table 2. E-Government services

Types	Definition
Government-to-Citizen (G2C)	It is an e-government service, from government to citizen in the form of offering valuable information and know-how's.
Citizen-to-Government (C2G)	It is an e-government service, offered for payment of bills and other valuable feedback from the citizen to government.
Government-to-Business (G2B)	It is an e-government service providing transactions and procurement facilities for government purchases and call for tenders.
Business -to-Government (B2G)	It is an e-government service providing communication, collaboration, transactions and procurement of goods and services for business initiatives.
Government-to-Employee (G2E)	It is an e-government initiative that will facilitate the management of the civil service and internal communication with governmental employees to encourage paperless office.
Government-to-Government (G2G)	It is an e-government initiative to provide the Government's departments or agencies cooperation and communication online. It includes internal exchange of information and commodities.
Government-to-Nonprofit (G2N)	It is an e-government initiative that provides information and communication from government to nonprofit organizations, political parties and social organizations, Legislature, etc.
Nonprofit-to-Government (N2G)	It is an e-government initiative that enable exchange of information and communication from non-profit organization to government organizations, political parties and social organizations, Legislature, etc.

4 Discussion: E-Government Adoption Factors

The adoption factors for e-government services should be thoroughly known before any adoption model is constructed. Many researchers have understood the initiatives that encourage the adoption of e-government services in different environments. These studies have shown that despite different environments having different characteristics, there are general initiatives that promote e-government adoption by ordinary citizens. However, it is worth mentioning that certain situations have unique factors which may either impend or aid the adoption of e-government services. In order to have a basic understanding of these varying factors, this study review's the adoption models that have been studied in different locations. The factors that influence the adoption of e-government websites are information quality, system quality and service quality. Information quality is concerned with the measure of the information that the system produces and delivers i.e., characteristics of information produced by e-government Web sites. Quality of information is believed to be the most salient factor for predicting customer decision-making behavior and user intention to use a particular system [30]. The fundamental dimensions of information quality are composed of five dimensions: accuracy, timeliness, relevance, understandability, and completeness [31–33].