Editors
Robert Stahlbock, Sven F. Crone
Stefan Lessmann

Associate Editors
Mahmoud Abou-Nasr, Hamid R. Arabnia
Philippe Lenca, Yanjun Li
Wolfram-M. Lippe, Anthony Scime
Gary M. Weiss

WORLDCOMP’09
July 13-16, 2009
Las Vegas Nevada, USA
www.world-academy-of-science.org

©CSREA Press
This volume contains papers presented at The 2009 International Conference on Data Mining (DMIN'09). Their inclusion in this publication does not necessarily constitute endorsements by editors or by the publisher.

Copyright and Reprint Permission

Copying without a fee is permitted provided that the copies are not made or distributed for direct commercial advantage, and credit to source is given. Abstracting is permitted with credit to the source. Please contact the publisher for other copying, reprint, or republication permission.

Copyright © 2009 CSREA Press
Printed in the United States of America

CSREA Press
U. S. A.
Foreword

We are pleased to present this collection of papers submitted to the 5th International Conference on Data Mining 2009, DMIN'09 (www.dmin--2009.com), July 13-16, 2009, held annually at the Monte Carlo Resort, Las Vegas, Nevada, USA.

Data mining continues to attract innovative and influential contributions to both research and practice, across an ever increasing range of academic disciplines and application domains. DMIN conferences seek to acknowledge and facilitate excellence in research and applications in the area of data mining. To reflect the multi- and interdisciplinary nature of data mining, only few conferences are better suited to facilitate the exchange and development of novel ideas, open communication and networking amongst researchers and practitioners in different fields than the DMIN conferences held annually within WORLDCOMP, the largest annual gathering of researchers in computer science, computer engineering and applied computing. WORLDCOMP'09 assembles a spectrum of 22 affiliated research conferences, workshops, and symposiums into a coordinated research meeting. We hope that the 2009 International Conference on Data Mining will provide you with a forum to present your research in a professional environment, exchange ideas and network within a number of research areas that interact. DMIN conferences actively support students and beginning researchers from lesser developed countries, to allow a truly international networking and understanding. DMIN’09 has supported the research of students and researchers from lesser developed countries by funding registration and accommodation. The 2009 conference has provided an international and multicultural experience with contributions from 27 different countries. We consider the resulting diversity in attendees and the mixture of established and starting researchers as a particular advantage of an engaging conference format.

DMIN’09 attracted a large number of submissions of theoretical research papers as well as industrial reports and case studies on applications. The programme committee would like to thank all those who submitted papers for review. To reflect upon feedback from previous years we will continue to further extend the quality and rigor of the review process and the constructive feedback given within the reviews. To ensure a fair, objective and transparent review process all review criteria were published on the website. Papers were evaluated regarding their relevance to DMIN, originality, significance, information content, clarity, and soundness on an international level. Each aspect was objectively evaluated, with alternative aspects finding consideration for application papers. Each paper was refereed by at least two researchers in the topical area, taken the reviewers’ expertise and confidence into consideration, with many papers receiving three and up to five reviews. The review process was highly competitive. We are very grateful to the many colleagues who helped in organising the conference. In particular, we would like to thank the members of the DMIN'09 programme committee. Their continuing support has been essential to further the quality of accepted submissions and hence success of the conference. The DMIN'09 programme committee members are: Mahmoud Abou-Nasr (USA), Vasilis Aggelis (Greece), Plamen Angelov (UK), Lamine Aouad (Ireland), Mohsen Askari (Iran), Daniel Berrar (UK), Christos Bouras (Greece), Alina Campan (Romania), Stephen Chi Fai Chan (China), Frans Coenen (UK), Paulo Cortez (Portugal), Sven F. Crone (UK), Kevin Daimi (USA), Thanh-Nghi Do (Vietnam), Xu E (China), William Eberle (USA), Xiaohua Anna Feng (UK), Peter Geczy (Japan), Guo Gongde (China), Monte F. Hancock, Jr. (USA), Haibo He (USA), Thomas J. Heiman (USA), Beatriz de la Iglesia (UK), Masoud Jamei (UK), Radha Krishna Murthy Karuturi (Singapore), Waldemar Koczkodaj (Canada), Rikard König (Sweden), Nikolaos Kourentzes (UK), Stéphane Lallich (France), Chung-Hong Lee (Taiwan), Yue-Shi Lee (Taiwan), Philippe Lenca (France), Stefan Lessmann (Germany), Xiaoli Li (UK), Yanjun Li (USA), Wen-Yang Lin (Taiwan), Wolfram-M. Lippe (Germany), Honghai Liu (UK), P. K. Mahanti (Canada), D. H. Manjaiah (India), Guojun Mao (China), Jun Meng (China), Patrick Meyer (France), Frederick Moxley (USA), Maybin Muyeba (UK), Mohamed Nadif (France), Alberto Ochoa-Zezzatti
(Mexico), Witold Pedrycz (Canada), Vassilis Poulopoulos (Greece), R. Rajesh (India), KVSVPN Raju (India), Torsten Reiners (Germany), Zhang Sen (USA), Xuequn Shang (China), Robert Stahlbock (Germany), Ryszard Tadeusiewicz (Poland), Traian Marius Truta (USA), Andreea Vesean (Romania), Baoying Wang (USA), Gary M. Weiss (USA), Show-Jane Yen (Taiwan), Liu Ying (China), and Jacek Zurada (USA). Furthermore, we would like to thank the dedicated reviewers: Gregory Baramidze (USA), Cécile Favre (France), Cristian Figueroa (Chile), Sylvie Guillaume (France), Shan He (UK), Tzung-Pei Hong (Taiwan), Zhbin Huang (China), Hung-Yu Kao (Taiwan), Ming-Yen Lin (Taiwan), Hongbo Liu (China), Qi Liu (China), Haibo Tang (USA), and Yan Zhang (USA). We are also grateful to our publicity co-chairs Ashu Solo, Maverick Technologies America, Wilmington DE, USA, and Innar Liiv, Tallinn University of Technology, Tallinn, Estonia, for circulating information on the conference. Considering the increasing efforts of all towards the quality of the review process, the conference sessions and the social programme of DMIN'09 we are confident that you can look forward to participating and attending a leading and reputable international conference. It is a particular pleasure to provide four tutorials during the evenings of DMIN’09, which will be held by esteemed members of the data mining community: Nitesh V. Chawla, Peter Geczy, Asim Roy, and Dan Steinberg.

The DMIN'09 conference organisers are particularly thankful to a number of co-sponsors, without whom the conference would not have been possible. The Academic Co-Sponsors of this year’s conference include: United States Military Academy, Network Science Center, USA; Biomedical Cybernetics Laboratory, HST of Harvard University and MIT, USA; Argonne's Leadership Computing Facility of Argonne National Laboratory, USA; Functional Genomics Laboratory, University of Illinois at Urbana-Champaign, USA; Minnesota Supercomputing Institute, University of Minnesota, USA; Intelligent Data Exploration and Analysis Laboratory, University of Texas at Austin, Austin, Texas, USA; Harvard Statistics Department Genomics & Bioinformatics Laboratory, Harvard University, USA; Texas Advanced Computing Center, The University of Texas at Austin, Texas, USA; Center for the Bioinformatics and Computational Genomics, Georgia Institute of Technology, Atlanta, Georgia, USA; Bioinformatics & Computational Biology Program, George Mason University, Virginia, USA; Institute of Discrete Mathematics and Geometry, Vienna University of Technology, Austria; BioMedical Informatics & Bio-Imaging Laboratory, Georgia Institute of Technology and Emory University, Atlanta, Georgia, USA; Knowledge Management & Intelligent System Center (KMIS) of University of Siegen, Germany; National Institute for Health Research; Hawkeye Radiology Informatics, Department of Radiology, College of Medicine, University of Iowa, Iowa, USA; Institute for Informatics Problems of the Russian Academy of Sciences, Moscow, Russia; Medical Image HPC & Informatics Lab (MiHi Lab), University of Iowa, Iowa, USA; SECLAB of University of Naples Federico II, University of Naples Parthenope, and the Second University of Naples, Italy; The University of North Dakota, Grand Forks, North Dakota, USA; Intelligent Cyberspace Engineering Lab., ICEL, Texas A&M University (Com./Texas), USA; International Society of Intelligent Biological Medicine; and World Academy of Biomedical Sciences and Technologies.

Corporate Co-Sponsors, Co-Sponsors At-Large and Organizers include: A number of university faculty members and their staff (names appear below and also on the cover of the proceedings); World Academy of Science (www.world-academy-of-science.org/); Computer Science Research, Education, and Applications Press; European Commission; Salford Systems, USA; Element CXI, California, USA; SuperMicro Computer, Inc., San Jose, California, USA; High Performance Computing for Nanotechnology (HPCNano); The International Council on Medical and Care Compunetics; The UK Department for Business, Enterprise & Regulatory Reform, UK; VMW Solutions Ltd.; Scientific Technologies Corporation; HolP - Health without Boundaries; Hodges' Health; Bentham Science Publishers; and GridToday. In addition to the above, several publishers of computer science and computer engineering books and journals, chapters and/or task forces of computer science associations/organizations from 9 countries, and developers of high-performance machines and systems provided significant help in organizing the
conference. We are also grateful for the general co-sponsors and organisers including university faculty members and their staff from the Institute of Information Systems at Hamburg University, Germany (www.uni-hamburg.de/IWI), the Centre for Forecasting and Predictive Intelligence at Lancaster University Management School, UK (www.lums.lancs.ac.uk/ forecasting), the World Academy of Science (www.world-academy-of-science.org), CSREA Computer Science Research, Education, and Applications Press, and the Business Intelligence Laboratory, B I’S lab, Hamburg, Germany (www.bis-lab.com).

Most importantly, we wish to express our sincere gratitude and respect towards Professor Hamid R. Arabnia, General Chair of all WORLDCOMP conferences, for his excellent and tireless support, organisation and coordination of all affiliated events. Without his exemplary and professional effort none of these events would be possible!

Thank you all for your contribution to DMIN’09! We hope that you will experience a stimulating conference with many opportunities for future contacts, research and applications.

Robert Stahlbock  Sven F. Crone
Stefan Lessmann

DMIN’09 General Conference Chair  DMIN’09 Conference Programme Co-Chairs
Contents

SESSION: A TOUR OF ADVANCED DATA MINING METHODOLOGIES

A Tour of Advanced Data Mining Methodologies: The CART Decision Tree 3
Dan Steinberg

SESSION: ASSOCIATION RULE MINING

Pruning for Extracting Class Association Rules Without Candidate Generation 11
Emna Bahri, Stephane Lallich

Mining Frequent Itemsets by Transaction Decomposition with Itemset Clustering 18
I-En Liao, Ke-Chung Lin, Hong-Bin Chen

Linked List Based High Utility Itemsets Mining Using Pattern Growth Method 25
Fan Wu, Ya-Han Hu, Kai-Chung Pai

SESSION: BUSINESS INTELLIGENCE

OLAP For Multicriteria Maintenance Scheduling 35
Walter Cai, David C. Anastasiu, Mingji Xia, Byron J. Gao

Using Social Ties to Predict Missing Customer Information 42
Stamatis Stefanakos

Advanced Implementation Techniques for Scientific Data Warehouses 48
Payyavula Gangadhar, Ravulapalli Lakshmi Tulasi

Analyzing Student Retention with Data Mining 55
Kevin Daimi, Ruth Miller

Automobile Insurance Knowledge Mining 61
Sulaiman Al-Hudhaif

SESSION: REAL-WORLD DATA MINING APPLICATIONS, CHALLENGES, AND PERSPECTIVES

Relevant Feature Selection and Generation in High Dimensional Haptic-based Biometric Data 71
Nizar Sakr, Fawaz Alsulaiman, Julio Valdes, Abdulmotaleb El Saddik, Nicolas Georganas

A Prime Number-Based Method for Interactive Frequent Pattern Mining 78
<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRC: Multi Relational Clustering approach</td>
<td>84</td>
</tr>
<tr>
<td>Majid Rastegar-Mojarad, Behrouz Minaei-Bidgoli</td>
<td></td>
</tr>
<tr>
<td>Concurrent Agent-enabled Extraction of Computational Fluid Dynamics (CFD) Features in Simulation</td>
<td>90</td>
</tr>
<tr>
<td>Clifton Mortensen, Robert Woodley, Steve Gorrell</td>
<td></td>
</tr>
<tr>
<td>Fish or Shark - Data Mining Online Poker</td>
<td>97</td>
</tr>
<tr>
<td>Ulf Johansson, Cecilia Sonstrood</td>
<td></td>
</tr>
<tr>
<td>An Efficient Clustering Algorithm based on Sorting and Binary Splitting</td>
<td>104</td>
</tr>
<tr>
<td>Taewan Ryu, Jae Soo Yoo, Michael Allen Bickel</td>
<td></td>
</tr>
<tr>
<td>Perturbation Scheme for Online Incremental Learning of Features for Face Recognition</td>
<td>110</td>
</tr>
<tr>
<td>R.K. Agrawal, Ashish Chaudhary</td>
<td></td>
</tr>
<tr>
<td>A Linear Integer Programming Approach to Objective Aware Feature Selection</td>
<td>117</td>
</tr>
<tr>
<td>Guangzhi Qu, Hui Wu, Tao Xia</td>
<td></td>
</tr>
<tr>
<td>Data Mining in the Real World: Experiences, Challenges, and Recommendations</td>
<td>124</td>
</tr>
<tr>
<td>Gary Weiss</td>
<td></td>
</tr>
<tr>
<td>Privacy Preserving Sharing of Data</td>
<td>131</td>
</tr>
<tr>
<td>Sreenivasa Rao, S. Ram Prasad Reddy, KV Ramana, V. Valli Kumari, KVSVN Raju</td>
<td></td>
</tr>
<tr>
<td>A Decision Support System Based on Data Mining for Pediatric Cardiology Diagnosis</td>
<td>138</td>
</tr>
<tr>
<td>Paulo Adeodato, Tarcísio Gurgel, Sandra Mattos</td>
<td></td>
</tr>
<tr>
<td>Name Entity Recognition and Classification in Medical Text Documents</td>
<td>144</td>
</tr>
<tr>
<td>Yinghao Huang, Naeem Seliya, Yi Lu Murphey, Roy B. Friedenthal</td>
<td></td>
</tr>
<tr>
<td>Distributed Frequent Pattern Mining Using Time-Slice Method</td>
<td>151</td>
</tr>
<tr>
<td>Fan Wu, Ya-Han Hu, Che-Wei Yang</td>
<td></td>
</tr>
<tr>
<td>A Novel and Efficient Distributed Data Mining Algorithm Based on Frequent Pattern-Tree</td>
<td>158</td>
</tr>
<tr>
<td>Fan Wu, Ya-Han Hu, Tz-Ke Wu</td>
<td></td>
</tr>
<tr>
<td>Factors Which Influence the Recovery of Alcohol Addicts: A Second Follow Up Study</td>
<td>165</td>
</tr>
<tr>
<td>Kathryn Burn-Thornton, Tim Burman</td>
<td></td>
</tr>
</tbody>
</table>
Supervised DAG Based Data Mining Model for DNA Sequence Analysis and Pattern Discovery
Shams-ul-haq Syed, Nadeem Muhammad

Performance Monitoring and Evaluation of Software Developers in an Information Technology Company using Data Mining Techniques
Chandrani Singh, Arpita Gopal

Are Decision Trees Always Greener on the Open (Source) Side of the Fence?
Samuel Moore, Daniel D'Addario, James Kurinskas, Gary Weiss

Action Selection in Customer Value Optimization: An Approach Based on Covariate-Dependent Markov Decision Processes
Angi Roesch, Harald Schmidbauer

A Language Modeling Approach for the Classification of Music Pieces
Gonçalo Marques, Thibault Langlois

SESSION: DATA MINING FOR TIME SERIES DATA - FORECASTING, CLASSIFICATION AND CLUSTERING

Dynamic Data Mining: A Novel Data Mining Process Model
Xiong Deng, Yike Guo, Moustafa Ghanem

ATM's Daily Cash Money Demand Forecasting with Recurrent Neural Networks
Mahmoud Abou-Nasr

Incorporating Conditional Probability Functions in Time-Dependent Bayesian Networks
Dung Lam, Cheryl Martin

Mining Sequential Episodes from Segmentation of Multivariate Time Series
Li Wan, Jianxin Liao, Xiaomin Zhu

An Algorithm For Efficiently Clustering High-dimensional Data Streams
Guojun Mao, Xialing Yang

Forecasting Seasonal Time Series with Multilayer Perceptrons - An Empirical Evaluation of Input Vector Specifications for Deterministic Seasonality
Sven F. Crone, Nikolaos Kourentzes

A Fuzzy Index Structure for Multi-Dimensional Data
Yong Shi
### SESSION: DATA PRE-PROCESSING

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fused Multi-modal Deduplication</td>
<td>253</td>
</tr>
<tr>
<td>Sabra Dinerstein, Christophe Giraud-Carrier, Jared Dinerstein, Parris K. Egbert</td>
<td></td>
</tr>
<tr>
<td>A Combinatorial Fusion Method for Feature Construction</td>
<td>260</td>
</tr>
<tr>
<td>Ye Tian, Gary Weiss, D. Frank Hsu, Qiang Ma</td>
<td></td>
</tr>
<tr>
<td>Influence of Noisy and High-Dimensional Data on Semi-Supervised Learning</td>
<td>267</td>
</tr>
<tr>
<td>Tianya Hou, Hyunjung Shin</td>
<td></td>
</tr>
<tr>
<td>Efficient Record Linkage using a Double Embedding Scheme</td>
<td>274</td>
</tr>
<tr>
<td>Noha Adly</td>
<td></td>
</tr>
<tr>
<td>Feature Extraction for Graph Datasets</td>
<td>282</td>
</tr>
<tr>
<td>Gideon Dror</td>
<td></td>
</tr>
</tbody>
</table>

### SESSION: PREDICTIVE MODELLING

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-Fold Cross Validation for Error Rate Estimate in Support Vector Machines</td>
<td>291</td>
</tr>
<tr>
<td>Davide Anguita, Alessandro Ghio, Sandro Ridella, Dario Sterpi</td>
<td></td>
</tr>
<tr>
<td>Rule-Based Linear Regression Machine</td>
<td>298</td>
</tr>
<tr>
<td>Olutayo Oladunni</td>
<td></td>
</tr>
<tr>
<td>Understanding Support Vector Machine Classifications via a Recommender System-Like Approach</td>
<td>305</td>
</tr>
<tr>
<td>David Barbella, Sami Benzaid, Janara Christensen, Bret Jackson, X. Victor Qin, David Musicant</td>
<td></td>
</tr>
<tr>
<td>View of Boosting as a Search for Randomness Deficiencies</td>
<td>312</td>
</tr>
<tr>
<td>Daniel Burfoot, Yasuo Kuniyoshi</td>
<td></td>
</tr>
<tr>
<td>A Probabilistic Model of Pattern Recognition on Abstract Data</td>
<td>319</td>
</tr>
<tr>
<td>Chun-Hung Tzeng</td>
<td></td>
</tr>
<tr>
<td>On Adaboost and Optimal Betting Strategies</td>
<td>326</td>
</tr>
</tbody>
</table>
Pasquale Malacaria, Fabrizio Smeraldi

An Experimental Study on a New Ensemble Method using Robust and Order Statistics 333
Faisal Zaman, Hideo Hirose

Large Experiment and Evaluation Tool for WEKA Classifiers 340
Dustin Baumgartner, Gursel Serpen

Formal Model For Fault Recognition In Modern Telecommunication Networks 347
Jacques-H Bellec, Tahar-M Kechadi

Evaluating Algorithms for Concept Description 354
Cecilia Sonstro, Ulf Johansson, Tuve Lofstrom

Greedy Learning: Using Advantages of Distribution Functions to Improve Generalization of ANNs 361
Kristina Davoian, Wolfram-M. Lippe

Generation of Weak Models in Stochastic Discrimination 368
Iryna Skrypanyk

Regression based Incremental Learning through Cluster Analysis of Temporal data 375
Syed Zakir Ali, Nagabhushan P, Pradeep Kumar R

Study of Dhaka Stock Exchange-Efficient Market Hypothesis 382
Jarka Arefin, Rashedur M Rahman

Identifying Subscribers with Multi-Split Fuzzy Decision Trees 389
Jaya Suma G, Shashi M

On The Usage of Data Mining as a Descriptive and Predictive Tool for Cancer Management in Jordan: A Scenario 396
Asem Omari, Issa Hweidi

SESSION: PRIVACY PRESERVING DATA MINING

P-Sensitive K-Anonymity for Social Networks 403
Roy Ford, Traian Marius Truta, Alina Campan

APHID: A Practical Architecture for High-Performance, Privacy-Preserving Data Mining 410
Jimmy Secretan, Anna Koufakou, Michael Georgiopoulos

High Speed Data Perturbation Methods for Privacy Preserving Data Mining 417
SESSION: TEXT AND WEB MINING

Personality Based Latent Friendship Mining 427
_Fan Wang, Yuan Hong, Wenbin Zhang, Gagan Agrawal_

Improved k-NN Algorithm for Text Classification 434
_Muhammed Miah_

SVM Fuzzy Hierarchical Document Categorization and Retrieval Method 441
_Taoufik Guernine, Kacem Zeroual_

Evaluation of Structured and Un-Structured Document Classification Techniques 448
_Saifullah Azim, Sohail Asghar_

Towards Online Personalized Foreseeing System by New Approach through Web Usage Mining 458
_Mehrdad Jalali, Norwati Mustapha, MD Nasir Sulaiman, Ali Mamat_

Algebraic Algorithms to Solve Name Disambiguation Problem 468
_Ingyu Lee, Byung-Won On, Seong No Yoon_

Improving Cohesiveness of Text Document Clusters 475
_Gaurav Ruhela_

Impact of a New Attribute Extraction Algorithm on Web Page Classification 481
_Göksel Biricik, Banu Diri_

On the Ranking of Text Documents from Large Corpuses 486
_Houssain Kettani, Greg Newby_

EmotiBlog: an Annotation Scheme for Emotion Detection and Analysis in Non-traditional Textual Genres 491
_Ester Boldrini, Alexandra Balahur, Patricio Martínez-Barco, Andres Montoyo_

Learning a Hyperplane using Genetic Algorithm for Sentiment Classification in Text 498
_Anjum Gupta_

SESSION: UNSUPERVISED DATA MINING

Outlier Detection for Large High-Dimensional Categorical Data using Non-Derivable and Non-Almost-Derivable Sets 505
SESSION: DATA MINING IN THE SOCIAL AND BEHAVIORAL SCIENCES

The Use of Logistic Regression Analyses and Data Classification Mining to Examine Variables Predictive of Long-term Healthcare Staff Giving Cessation Advice

Celia Watt, Jill Lassiter, Douglas Scheidt

Testing Terrorism Using Iterative Expert Data Mining

Gregg R. Murray, Lance Y. Hunter, Anthony Scime

SESSION: LATE PAPERS

Mining Spreadsheet Complexity Data to Classify End User Developers

Stephen Hole, Duncan McPhee, Alex Lohfink
Extreme Learning Machine Classifier Capabilities in Solving Multicategory Disease Classification Problems
Emad El-Sebakhy, Tarek Sheltami, Asparouhov Asparouhv, Krassimir Latinski, Zeehasham Rasheed

Parallel Synchronization of View Definitions Based On Clustering Techniques
Ines HILALI JAGHDAM, Jalel AKAICHI

The Analysis of the Relationships Between the Real Vehicle Information and the Fuel Consumption
Jongwoo Choi, Daesub Yoon, Kyoungho Kim, Hyunsuk Kim

A Comparative Study of Simulated Annealing and Variable Neighborhood Search for the Geographic Clustering Problem
Beatriz Bernábe Loranca, Jose Rosales Espinosa, Maria Auxilio Osorio Lama, Javier Ramírez Rodríguez, Ricardo García Aceves