

**PROCEEDINGS OF
THE 2009 INTERNATIONAL CONFERENCE ON
DATA MINING**

DMIN 2009

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**



WORLD COMP'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
ISBN: 1-60132-099-X
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 Kourntzes (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 Pouloupoulos (Greece), R. Rajesh (India), KVSVN Raju (India), Torsten Reiners (Germany), Zhang Sen (USA), Xuequn Shang (China), Robert Stahlbock (Germany), Ryszard Tadeusiewicz (Poland), Traian Marius Truta (USA), Andreea Vescan (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), Zhibin 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; HoIP - 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 I³S 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

Mohammad Nadimi-Shahraki, Norwati Mustapha, Md Nasir Sulaiman, Ali Mamat

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

Supervised DAG Based Data Mining Model for DNA Sequence Analysis and Pattern Discovery	171
<i>Shams-ul-haq Syed, Nadeem Muhammad</i>	
Performance Monitoring and Evaluation of Software Developers in an Information Technology Company using Data Mining Techniques	178
<i>Chandrani Singh, Arpita Gopal</i>	
Are Decision Trees Always Greener on the Open (Source) Side of the Fence?	185
<i>Samuel Moore, Daniel D'Addario, James Kurinkas, Gary Weiss</i>	
Action Selection in Customer Value Optimization: An Approach Based on Covariate-Dependent Markov Decision Processes	189
<i>Angi Roesch, Harald Schmidbauer</i>	
A Language Modeling Approach for the Classification of Music Pieces	193
<i>Gonçalo Marques, Thibault Langlois</i>	
SESSION: DATA MINING FOR TIME SERIES DATA - FORECASTING, CLASSIFICATION AND CLUSTERING	
Dynamic Data Mining: A Novel Data Mining Process Model	199
<i>Xiong Deng, Yike Guo, Moustafa Ghanem</i>	
ATM's Daily Cash Money Demand Forecasting with Recurrent Neural Networks	207
<i>Mahmoud Abou-Nasr</i>	
Incorporating Conditional Probability Functions in Time-Dependent Bayesian Networks	214
<i>Dung Lam, Cheryl Martin</i>	
Mining Sequential Episodes from Segmentation of Multivariate Time Series	221
<i>Li Wan, Jianxin Liao, Xiaomin Zhu</i>	
An Algorithm For Efficiently Clustering High-dimensional Data Streams	228
<i>Guojun Mao, Xialing Yang</i>	
Forecasting Seasonal Time Series with Multilayer Perceptrons - An Empirical Evaluation of Input Vector Specifications for Deterministic Seasonality	232
<i>Sven F. Crone, Nikolaos Kourentzes</i>	
A Fuzzy Index Structure for Multi-Dimensional Data	239
<i>Yong Shi</i>	

Network-wide Analysis Using Spatio-temporal Association Rules Mining 243
Weisong He, Guangmin Hu

Abnormal Process State Detection by Cluster Center Point Monitoring in BWR Nuclear Power Plant 247
Jaakko Talonen, Miki Sirola

SESSION: DATA PRE-PROCESSING

Fused Multi-modal Deduplication 253
Sabra Dinerstein, Christophe Giraud-Carrier, Jared Dinerstein, Parris K. Egbert

A Combinatorial Fusion Method for Feature Construction 260
Ye Tian, Gary Weiss, D. Frank Hsu, Qiang Ma

Influence of Noisy and High-Dimensional Data on Semi-Supervised Learning 267
Tianya Hou, Hyunjung Shin

Efficient Record Linkage using a Double Embedding Scheme 274
Noha Adly

Feature Extraction for Graph Datasets 282
Gideon Dror

SESSION: PREDICTIVE MODELLING

K-Fold Cross Validation for Error Rate Estimate in Support Vector Machines 291
Davide Anguita, Alessandro Ghio, Sandro Ridella, Dario Sterpi

Rule-Based Linear Regression Machine 298
Olutayo Oladunni

Understanding Support Vector Machine Classifications via a Recommender System-Like Approach 305
David Barbella, Sami Benzaid, Janara Christensen, Bret Jackson, X. Victor Qin, David Musicant

View of Boosting as a Search for Randomness Deficiencies 312
Daniel Burfoot, Yasuo Kuniyoshi

A Probabilistic Model of Pattern Recognition on Abstract Data 319
Chun-Hung Tzeng

On Adaboost and Optimal Betting Strategies 326

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 Sonstrod, 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 Skrypnyk

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
<i>Fan Wang, Yuan Hong, Wenbin Zhang, Gagan Agrawal</i>	
Improved k-NN Algorithm for Text Classification	434
<i>Muhammed Miah</i>	
SVM Fuzzy Hierarchical Document Categorization and Retrieval Method	441
<i>Taoufik Guernine, Kacem Zeroual</i>	
Evaluation of Structured and Un-Structured Document Classification Techniques	448
<i>Saifullah Azim, Sohail Asghar</i>	
Towards Online Personalized Foreseeing System by New Approach through Web Usage Mining	458
<i>Mehrdad Jalali, Norwati Mustapha, MD Nasir Sulaiman, Ali Mamat</i>	
Algebraic Algorithms to Solve Name Disambiguation Problem	468
<i>Ingyu Lee, Byung-Won On, Seong No Yoon</i>	
Improving Cohesiveness of Text Document Clusters	475
<i>Gaurav Ruhela</i>	
Impact of a New Attribute Extraction Algorithm on Web Page Classification	481
<i>Göksel Biricik, Banu Diri</i>	
On the Ranking of Text Documents from Large Corporuses	486
<i>Houssain Kettani, Greg Newby</i>	
EmotiBlog: an Annotation Scheme for Emotion Detection and Analysis in Non-traditional Textual Genres	491
<i>Ester Boldrini, Alexandra Balahur, Patricio Martínez-Barco, Andres Montoyo</i>	
Learning a Hyperplane using Genetic Algorithm for Sentiment Classification in Text	498
<i>Anjum Gupta</i>	

SESSION: UNSUPERVISED DATA MINING

Outlier Detection for Large High-Dimensional Categorical Data using Non-Derivable and Non-Almost-Derivable Sets	505
--	------------

Anna Koufakou, Jimmy Secretan, Michelle Fox, Gary Gramajo, Georgios C. Anagnostopoulos, Michael Georgiopoulos

A Sparse Coding Based Similarity Measure 512

Sebastian Klenk, Gunther Heidemann

An Enhanced Density Based Spatial clustering of Applications with Noise 517

Ahmed Fahim, Gunter Saake, Abdel-badeeh Salem, Fawzy Torkey, Mohamed Ramadan

GCLOD : A Clustering Algorithm for Improved Intra-cluster Similarity and Efficient Local Outliers Detection 524

Abdul Aleem, Reena Srivastava, Anil Kumar Singh, M. M. Gore

A Graph-based Similarity Metric and Validity Indices for Clustering Non-numeric and Unstructured Data 531

Chuan Zhao, Krishnamoorthy Sivakumar

Comparison of Standard and Optimized K-means in SQL 538

Nedunchelian R, Muthucumarasamy R, Saranathan E

Opportunistic Consensus Clustering 543

Arko Banerjee, Arun K Pujari

Learning a Similarity Measure to Objectively Evaluate Image Segmentation Quality 550

Anjum Gupta

Improving the Results of Partial Match 555

Aikaterini Krotopoulou

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 561

Celia Watt, Jill Lassiter, Douglas Scheidt

Testing Terrorism Using Iterative Expert Data Mining 565

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

SESSION: LATE PAPERS

Mining Spreadsheet Complexity Data to Classify End User Developers 573

Stephen Hole, Duncan McPhee, Alex Lohfink

Extreme Learning Machine Classifier Capabilities in Solving Multicategory Disease Classification Problems 580

Emad El-Sebakhy, Tarek Sheltami, Asparouhov Asparouhov, Krassimir Latinski, Zeehasham Rasheed

Parallel Synchronization of View Definitions Based On Clustering Techniques 586

Ines HILALI JAGHDAM, Jalel AKAICHI

The Analysis of the Relationships Between the Real Vehicle Informaion and the Fuel Consumption 591

Jongwoo Choi, Daesub Yoon, Kyoungho Kim, Hyunsuk Kim

A Comparative Study of Simulated Annealing and Variable Neighborhood Search for the Geographic Clustering Problem 595

Beatriz Bernábe Loranca, Jose Rosales Espinosa, Maria Auxilio Osorio Lama, Javier Ramírez Rodríguez, Ricardo García Aceves