

**PROCEEDINGS OF
THE 2010 INTERNATIONAL CONFERENCE ON
DATA MINING**

DMIN 2010

Editors

**Robert Stahlbock
Sven F. Crone**

Associate Editors

**Mahmoud Abou-Nasr, Hamid R. Arabnia
Nikolaos Kourentzes, Philippe Lenca
Wolfram-M. Lippe, Gary M. Weiss**



WORLDCOMP'10

July 12-15, 2010

Las Vegas Nevada, USA

www.world-academy-of-science.org

©CSREA Press

This volume contains papers presented at The 2010 International Conference on Data Mining (DMIN'10). 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 © 2010 CSREA Press
ISBN: 1-60132-138-4
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 6th International Conference on Data Mining 2010, DMIN'10 (www.dmin-2010.com), July 12-15, 2010, 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 a wide 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. DMIN conferences are held annually within WORLDCOMP, the largest annual gathering of researchers in computer science, computer engineering and applied computing. WORLDCOMP'10 assembles a spectrum of 22 affiliated research conferences, workshops, and symposiums into a coordinated research meeting. As such, DMIN seeks to reflect the multi- and interdisciplinary nature of data mining and to facilitate the exchange and development of novel ideas, open communication and networking amongst researchers and practitioners in different research domains. We hope that the 2010 International Conference on Data Mining will provide you with a forum to present your research in a professional environment, exchange ideas, and network and interact across research areas. DMIN conferences actively support students and beginning researchers from lesser developed countries by funding registration and accommodation, in order to allow for a truly international networking and understanding. The 2010 conference has provided an international and multicultural experience with contributions from 25 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'10 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 strive 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 most of the 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 organizing the conference. In particular, we would like to thank the members of the DMIN'10 programme committee. Their continuing support has been essential to further improve the quality of accepted submissions and the resulting success of the conference. The DMIN'10 programme committee members are (in alphabetical order): Mahmoud Abou-Nasr (USA), Rayner Alfred (Malaysia), Plamen Angelov (UK), Lamine Aouad (Ireland), Daniel Berrar (UK), Alina Campan (USA), Dongsheng Che (USA), Paulo Cortez (Portugal), Sven F. Crone (UK), Kevin Daimi (USA), Christian Dawson (UK), Thanh-Nghi Do (Vietnam), William Eberle (USA), Cécile Favre (France), Mengling Feng (Singapore), Cristian Figueroa (Chile), Peter Geczy (Japan), Reza Ghaemi (Malaysia), Sylvie Guillaume (France), Haibo He (USA), Thomas J. Heiman (USA), Tzung-Pei Hong (Taiwan), Majid Vafaei Jahan (Iran), Mehrdad Jalali (Malaysia), Ulf Johansson (Sweden), Hung-Yu Kao (Taiwan), Rikard König (Sweden), Nikolaos Kourentzes (UK), Chung-Hong Lee (Taiwan), Yue-Shi Lee (Taiwan), Philippe Lenca (France), Chuan Li (China), Ming-Yen Lin (Taiwan), Wen-Yang Lin (Taiwan), Wolfram-M. Lippe (Germany), Qi Liu (China), P. K. Mahanti (Canada), Guojun Mao (China), Ken McGarry (UK), Patrick Meyer (France), Maybin Muyeba (UK), Mohamed Nadif (France), Mohammad-Hosseion Nadimi-Shahraki (Malaysia), Alberto Ochoa-Zezzatti (Mexico), Ping-Feng Pai (Taiwan), Vassilis Pouloupoulos (Greece), Guangzhi Qu (USA), Torsten Reiners (Germany),

Abdel-badeeh Salem (Egypt), Zhang Sen (USA), Xuequn Shang (China), Robert Stahlbock (Germany), Sang Suh (USA), Shiliang Sun (China), Ying Tan (China), Traian Marius Truta (USA), Andreea Vescan (Romania), Baoying Wang (USA), Simon Wang (USA), Yue Wang (Singapore), Gary M. Weiss (USA), Dwi H Widyantoro (Indonesia), Show-Jane Yen (Taiwan), Yan Zhang (USA), Songfeng Zheng (USA), and Jacek Zurada (USA). We are grateful to our publicity co-chair Ashu Solo, Maverick Technologies America, Wilmington DE, USA, 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'10 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 two data mining oriented tutorials during the evenings of DMIN'10 presented by Vladimir Cherkassky and Peter Geczy who are esteemed members of the data mining community.

The DMIN'10 conference organizers are also thankful to a number of co-sponsors, without whom the conference would not have been possible. The Academic and Technical Co-Sponsors of this year's conference include: The Berkeley Initiative in Soft Computing (BISC), University of California, Berkeley, USA; Collaboratory for Advanced Computing and Simulations (CACs), University of Southern California, USA; Intelligent Data Exploration & Analysis Lab., University of Texas at Austin, Texas, USA; Harvard Statistics Department Genomics & Bioinformatics Lab., Harvard University, Massachusetts, USA; BioMedical Informatics & Bio-Imaging Lab., Georgia Institute of Technology & Emory University, Georgia, USA; Hawkeye Radiology Informatics, Department of Radiology, College of Medicine, University of Iowa, USA; Minnesota Supercomputing Institute, University of Minnesota, USA; Center for the Bioinformatics and Computational Genomics, Georgia Institute of Technology, USA; Medical Image HPC & Informatics Lab. (MiHi Lab), University of Iowa, USA; University of North Dakota, USA; NDSU-CIIT Green Computing & Communications Lab., North Dakota State University, USA; Knowledge Management & Intelligent System Center (KMIS) of University of Siegen, Germany; UMIT, Institute of Bioinformatics and Translational Research, Austria; SECLAB of University of Naples Federico II, University of Naples Parthenope, & Second University of Naples, Italy; National Institute for Health Research; World Academy of Biomedical Sciences and Technologies; High Performance Computing for Nanotechnology (HPCNano); Supercomputer Software Department (SSD), Institute of Computational Mathematics & Mathematical Geophysics, Russian Academy of Sciences; Int'l Council on Medical & Care Compunetics; The UK Department for Business, Innovation and Skills, UK; VMW Solution Ltd.; Scientific Technologies Corporation; HoIP - Health without Boundaries; Space for Earth Foundation; and Manjrasoft (Cloud Computing Technology company), Melbourne, Australia.

We are also grateful for the general co-sponsors and organisers including university faculty members 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'10! We hope that you will experience a stimulating conference with many opportunities for future contacts, research and applications.

Robert Stahlbock
DMIN'10 General Conference Chair

Sven F. Crone
DMIN'10 Conference Programme Chair

Contents

SESSION: DATA PRE-PROCESSING AND SAMPLING

SIPPA for Biometric Data Reconstruction	3
<i>Bon Sy</i>	
An Effective Entity Resolution Method	10
<i>Chuan Zhao, Krishnamoorthy Sivakumar</i>	
Cluster Infiltration in Fraud and Money Laundering Prediction	17
<i>Felipe Cardona, Gustavo Garcia, Juan Buritica</i>	
An Approach for Learning from Small and Unbalanced Data Sets Using Gaussian Noise During Artificial Neural Network Training	23
<i>Icamaan Viegas, Paulo Adeodato</i>	

SESSION: WEB AND TEXT MINING

Enhanced Named Entity Extraction via Error-Driven Aggregation	31
<i>Tracy Lemmond, Nathan Perry, Joseph Guensche, John Nitao, Ronald Glaser, Paul Kidwell, William Hanley</i>	
Finding Topic-specific Strings in Text Categorization and Opinion Mining Contexts	38
<i>Remi Lavalley, Chloe Clavel, Marc El-Beze, Patrice Bellot</i>	
Using and Parsing Wikipedia Articles For Text Classification	45
<i>Anjum Gupta, Craig Martell</i>	
A Hybrid Filter Car Recommender System	51
<i>Jayalakshmi Jagadeesan, Robert Chun</i>	
Classifying Independent Medical Examination Reports using SOM Networks	58
<i>Yinghao Huang, Naeem Seliya, Yi Murphey, Roy Friedenthal</i>	
Tuning Topical Query Classification in Large Search Engines to Understand Hidden User Information Need	65
<i>Abdulbaghi Ghaderzadeh, Behrouz Minaie Bidgoli</i>	
A Framework for Social Spam Detection Based on Relational Bayes Classifier	71
<i>Ahmet Aycan Atak, Sule Gunduz Oguducu</i>	

Detection of MMORPG Misconducts Based on Action Frequencies, Types and Time-Intervals 78
Ruck Thawonmas, Yoshitaka Kashifuji

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

Preliminary Results of Ranking Political Figures Using Naive Bayes Text Classification 85
James Ryder, Sen Zhang

A BI 2.0 Application Architecture for Healthcare Data Mining Services in the Cloud 92
Joseph M Woodside

Performance Evaluation of Intrusion Detection System Using Optimal Decision Tree SVM 98
Manju Bala, Namita Aggrawal, R. K. Agrawal

**Learner Interaction Monitoring System (LiMS): Capturing the Behaviors of Online Learners 106
and Evaluating Online Training Courses**
Peter Sorenson, Leah P. Macfadyen

**Elaboration of Applied Technologies for Modelling Connected Geomechanical, Geofiltration 112
and Geodynamic Processes in Rock's Massif**
Michael Zhuravkov, Aleg Kanavalau

A New Technique For Feature Selection And Cluster Center Initialization 119
Dharmveer Singh Rajoot, Pramod K. Singh, Mahua Bhattacharya

Examining Class Dropping with Data Mining 126
Kevin Daimi, Lu Wang

Predicting Protein Secondary Structure: A Recurrent Neural Network Approach 133
Mahmoud Abou-Nasr

Using Data Mining to Analyze Patient Discharge Data for an Urban Hospital 139
Xiaochun Jiang, Xiuli Qu, Lauren Davis

Application of Data Mining for Breast Cancer Survivability 145
Fatemeh Hosseinkhah, Hassan Ashktorab, Mohammad Owrang O.

Winning Baseball Through Data Mining 151
Gregory Swan , Anthony Scime

An Application of Clustering Techniques to Urban Studies 158
Zahra Ferdowsi, Raffaella Settini, Daniela Raicu

Peeling Decision Tree Based on Fisher's Linear Discriminant Analysis	165
<i>Jie Ouyang, Nilesh Patel</i>	
Hybrid PSO Algorithms for Dynamic Clustering	171
<i>Andrea Villagra, Daniel Pandolfi, Guillermo Leguizamon</i>	
Revenue Generation in Hospital Foundations: Neural Network versus Regression Model Recommendations	181
<i>Mary Malliaris, Maria Pappas</i>	
Real-time Automatic Building Identification	187
<i>Robert Woodley, Warren Noll, Joseph Barker</i>	
Association Analysis of Semi-structured Data for Discrimination Discovery in Business	193
<i>Binh Luong Thanh, Franco Turini</i>	
WIBE: Mining Frequent Closed Patterns Without Candidate Maintenance in Microarray Dataset	200
<i>Miao Wang, Xuequn Shang, Jingni Diao, Zhanhuai Li</i>	
s-Tuple Inclusion - A New Method for Privacy Preserving Publication of Datasets	206
<i>Ram Prasad Reddy S, Valli Kumari V, VSVN Raju K</i>	
Knowledge Discovery in the Virtual Social Network Due to Common Knowledge of Proverbs	213
<i>Armando Mendes, Matthias Funk, Luis Cavique</i>	
Music Databases and Data Mining Approaches	220
<i>Satyasaivani Bommakanti, Shashi M, Nagalakshmi V, Vikram D</i>	
A Student-Oriented University Ontology	228
<i>Annemie Vorstermans, Pavol Tanuska, Werner Verschelde, Michal Kebisek</i>	
SESSION: PREDICTIVE MODELLING	
SMO-Style Algorithms for Learning Using Privileged Information	235
<i>Dmitry Pechyony, Rauf Izmailov, Akshay Vashist, Vladimir Vapnik</i>	
Learning Ensemble Models on Categorized Datasets	242
<i>Xiong Deng, Yike Guo, Moustafa Ghanem</i>	
Solving Classification Problems with the NWEA-evolved ANNs	249
<i>Kristina Davoian, Wolfram-M. Lippe</i>	

Bi-Directional Subspace Decomposition in Classification	256
<i>Iryna Skrypnyk</i>	
Simple Method for Interpretation of High-Dimensional Nonlinear SVM Classification Models	267
<i>Vladimir Cherkassky, Sauprik Dhar</i>	
Inference for Neural Network Predictive Models with Impulse Interventions	273
<i>Nikolaos Kourentzes, Sven F. Crone</i>	
Ensembles of Locally Linear Models: Application to Bankruptcy Prediction	280
<i>Laura Kainulainen, Qi Yu, Yoan Miche, Emil Eirola, Eric Séverin, Amaury Lendasse</i>	
An Approach for Time Series Forecasting by Simulating Stochastic Processes Through Time Lagged Feed-forward Neural Network	287
<i>Cristian Rodriguez Rivero, Julian Pucheta, Josef Baumgartner, Hector D. Patino, B. Kuchen</i>	
Feature Selection Using a Novel Swarm Intelligence Algorithm with Rough Sets	294
<i>Noorhaniza Wahid, Yuk Ying Chung, Wei-Chang Yeh, Guang Liu</i>	
A Border-Based Approach for Hiding Fuzzy Weighted Sensitive Itemsets	301
<i>Durga Toshniwal, Mridula Verma</i>	
High Dimensional Data Classification by the PolSOM Algorithm	308
<i>Lu Xu, T. W. S. Chow</i>	
Data Mining and Electronic Health Records: Selecting Optimal Clinical Treatments in Practice	313
<i>Casey Bennett, Thomas Doub</i>	
Fat Tailed Distribution of Neural Networks Forecasting	319
<i>Domingos S. P. Salazar, Maíra de O. Santos, Adrian L. Arnaud, Paulo J. L. Adeodato</i>	
SESSION: SEGMENTATION, CLUSTERING, ASSOCIATION	
Clustering Scatter Plots Using Data Depth Measures	327
<i>Zhanpan Zhang, Xinping Cui, Daniel Jeske, Xiaoxiao Li, Jonathan Braun, James Borneman</i>	
Novel Significance Weighting Schemes for Collaborative Filtering: Generating Improved Recommendations in Sparse Environments	334
<i>Mustansar Ali Ghazanfar, Adam Prugel-Bennett</i>	
SP&EPPF: A Novel Algorithm for Sequential Interval-based Pattern Mining	343
<i>Jieh-Shan Yeh, Chia-Hsien Ting</i>	

Interactive Clustering of Proteomic Data - a Comparison between Self-Organizing Map and Neural Gas	350
<i>Vemund Jakobsen, Terje Kristensen</i>	
Extracting Interesting Patterns of Hard CSPs	357
<i>Chendong Li, Jiayin Wang, Yichen Liu</i>	
Application of Clustering in a Hybrid Global Optimization Algorithm	364
<i>Ting-Yu Chen, Jyun Hao Huang</i>	
Extension of Hierarchical Information Acquirement by Association Rule Mining from Semi-Structured Data	370
<i>Ryosuke Saga, Chihiro Sugaya, Kodai Kitami</i>	
Discovering Latent Healthy Nutritional Dietary Patterns with Association Rule Mining and Constraint Handling Rules	375
<i>Chendong Li, Yichen Liu</i>	
A Novel Architecture for Hiding Sensitive Association Rules	380
<i>Muhammad Naeem, Sohail Asghar</i>	
Finding Sequential Patterns from Temporal Datasets	386
<i>Fokrul Alom Mazarbhuiya, Anjana Kakoti Mahanta, Hemanta K. Baruah</i>	
Using Unsupervised Machine Learning Methods in High-Throughput Screening	392
<i>Chérif Mballo, Vladimir Makarenkov</i>	
A Dimension-wise Algorithm for Multi-dimensional Data	396
<i>Yong Shi, Ali Tavakoli</i>	
 SESSION: STOCHASTIC RESONANCE, CLASSIFICATION, AND INDEXING	
Stochastic Resonance Noise Benefits in Markov Chain Monte Carlo Density Estimation	403
<i>Brandon Franzke, Bart Kosko</i>	
Document Indexing by Latent Dirichlet Allocation	409
<i>In-Chan Choi, Jae-Sung Lee</i>	
Novel Hybrid Method for Sentiment Classification of Movie Reviews	415
<i>Zied Kechaou, Ali Wali, Mohamed Benammar, Adel M. Alimi</i>	

