

**PROCEEDINGS OF  
THE 2011 INTERNATIONAL CONFERENCE ON  
DATA MINING**

# **DMIN 2011**

**Editor**

**Robert Stahlbock**

**Associate Editors**

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



***WORLD COMP'11***

July 18-21, 2011

Las Vegas Nevada, USA

[www.world-academy-of-science.org](http://www.world-academy-of-science.org)

©CSREA Press

This volume contains papers presented at The 2011 International Conference on Data Mining (DMIN'11). 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 © 2011 CSREA Press  
ISBN: 1-60132-168-6  
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 7<sup>th</sup> International Conference on Data Mining 2011, DMIN'11 ([www.dmin--2011.com](http://www.dmin--2011.com)), July 18-21, 2011, Monte Carlo Resort, Las Vegas, Nevada, USA.

Data mining attracts 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'11 assembles a spectrum of 22 affiliated research conferences, workshops, and symposiums into a coordinated research meeting. Each conference has its own program committee as well as referees and own indexed proceedings. Attendees have full access to all 22 conferences' sessions, tracks, and tutorials. 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 2011 International Conference on Data Mining will provide a forum for you 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 2011 conference has provided an international and multicultural experience with contributions from 28 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'11 attracted a large number of submissions of theoretical research papers as well as industrial reports and case studies on applications. The program 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'11 program committee. Their continuing support has been essential to further improve the quality of accepted submissions and the resulting success of the conference. The DMIN'11 program committee members are (in alphabetical order):

Mahmoud Abou-Nasr, Paulo Adeodato, Leonidas Anastasakis, Lamine M. Aouad, Jérôme Azé, Souhaib Ben Taieb, Khalid Benabdeslem, Daniel Berrar, Jane Margot Binner, Alina Campan, Pedro A. Castillo, Peng Chen, Christian Dawson, Paulo Cortez, Kevin Daimi, Qin Ding, William Eberle, Georgios Evangelidis, Mohammed Farquad, Mengling Feng, Philippe Fournier-Viger, Shunkai Fu, Peter Geczy, Giorgio Corani, Arun Gupta, Zahid Halim, Hongmei He, Thomas J. Heiman, Kenneth E. Hild II, Tzung-Pei Hong, Wei-Chiang Hong, Yo-Ping Huang, Catherine Huang, Danish Irfan, Mehrdad Jalali, Ulf Johansson, Rikard König, Amir Hossein Keyhanipour, Madjid Khalilian, Sebastian Klenk, Nikolaos Kourentzes, Terje Kristensen, Yue-Shi Lee, Jaewook Lee, Philippe Lenca, Chuan Li, Wen-Yang Lin, Tran Hoai Linh, Wolfram-M. Lippe, Jing Liu, Qi Liu, Weifeng Liu, Tanja Magoc, Jun Meng, José M. Merigo Lindahl, Gaolin Zheng Milledge, Mohamed Nadif, Alberto Ochoa-Zezzatti, David L. Olson, Parag C. Pendharkar,

Antonio Luigi Perrone, Hossein Peyvandi, Vassilis Plagianakos, Guangzhi Qu, Mekki Rachida, Rabie A. Ramadan, Robert G. Reynolds, Lotfi Ben Romdhane, Motaz Saad, Ram Prasad Reddy Sadi, Gerald Schaefer, Zhang Sen, Sabrina Senatore, Xuequn Shang, Victor Sheng, Yong Shi, Tamanna Siddiqui, Vijendra Singh, Lay-Ki Soon, Robert Stahlbock, Shiliang Sun, Sundaram Suresh, Ryszard Tadeusiewicz, Jaakko Talonen, Ying Tan, Traian Marius Truta, Shun-Hung Tsai, NB Venkateswarlu, Nicole Vincent, Baoying Wang, Simon Wang, Fan Wang, Chamont Wang, Xuewei Wang, Gary Weiss, Tianyi Wu, Zijiang Yang, Bingru Yang, Faisal Zaman, Yun Zhai, Yan Zhang, Defu Zhang, Songfeng Zheng, and Shang-Ming Zhou.

We would also like to thank our publicity co-chair Ashu M. G. Solo (Fellow of British Computer Society, Principal/R&D Engineer, Maverick Technologies America Inc., Intelligent Systems Instructor, Trailblazer Intelligent Systems, Inc.), 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 program of DMIN'11 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 data mining oriented invited talks and tutorials presented by the following esteemed members of the data mining community: Nitesh V. Chawla, Peter Geczy, Michael Mahoney and Gary M. Weiss. Furthermore, it is planned to publish a Special Issue on Real World Data Mining Applications in Springer's Annals of Information Systems. In 2010 we have published a Special Issue on Data Mining in that series.

The DMIN'11 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 (<http://www-bisc.cs.berkeley.edu/>); Biomedical Cybernetics Laboratory, HST of Harvard University and Massachusetts Institute of Technology (MIT), USA (<http://bcl.med.harvard.edu/>); Intelligent Data Exploration and Analysis Laboratory, University of Texas at Austin, Austin, Texas, USA (<http://www.ideal.ece.utexas.edu/>); Collaboratory for Advanced Computing and Simulations (CACs), University of Southern California, USA (<http://cacs.usc.edu/>); Minnesota Supercomputing Institute, University of Minnesota, USA (<http://www.msi.umn.edu/>); Knowledge Management & Intelligent System Center (KMIS) of University of Siegen, Germany (<http://www.kmis.uni-siegen.de/>); UMIT, Institute of Bioinformatics and Translational Research, Austria (<http://www.umat.at/page.cfm?vpath=departments/technik/bioinf&expanddiv=subDeptItem343>); BioMedical Informatics & Bio-Imaging Laboratory, Georgia Institute of Technology and Emory University, Atlanta, Georgia, USA (<http://www.bio-miblab.org/>); Hawkeye Radiology Informatics, Department of Radiology, College of Medicine, University of Iowa, Iowa, USA (<http://www.uiowa.edu/~hri/>); NDSU-CIIT Green Computing and Communications Laboratory, USA (<http://gcclab.org/>); High Performance Computing for Nanotechnology (HPCNano) (<http://www.hpcnano.org/>); Supercomputer Software Department (SSD), Institute of Computational Mathematics & Mathematical Geophysics, Russian Academy of Sciences (<http://ssd.ssc.ru/>); SECLAB - An inter-university research group (University of Naples Federico II, the University of Naples Parthenope, and the Second University of Naples, Italy, <http://www.seclab.unina.it/>); Medical Image HPC & Informatics Lab (MiHi Lab), University of Iowa, Iowa, USA (<http://www.uiowa.edu/mihpclub/>); International Society of Intelligent Biological Medicine (<http://www.isibm.org/>); World Academy of Biomedical Sciences and Technologies (<http://www.worldwabt.org/wabt/>); Intelligent Cyberspace Engineering Lab., ICEL, Texas A&M University (Com./Texas), USA; Model-Based Engineering Laboratory, University of North Dakota, North Dakota, USA; The International Council on Medical and Care Compunetics (<http://www.icmcc.org/>); The UK Department for Business, Enterprise & Regulatory Reform (<http://www.berr.gov.uk/>); Scientific Technologies Corporation (<http://www.stchome.com/>); HoIP - Health without Boundaries (<http://www.hoip.eu/>).

We are also grateful for the general co-sponsors and organizers including university faculty members from the Institute of Information Systems at Hamburg University, Germany ([www.uni-hamburg.de/IWI](http://www.uni-hamburg.de/IWI)), CSREA Computer Science Research, Education, and Applications Press, and the Business Intelligence Laboratory, B I<sup>3</sup>S lab, Hamburg, Germany ([www.bis-lab.com](http://www.bis-lab.com)).

Furthermore, we want to thank all members of the steering committee of Worldcomp 2011: Selim Aissi (Intel Corp., USA), Ruzena Bajcsy (Univ. of California, Berkeley, USA), Hyunseung Choo (Sungkyunkwan Univ., South Korea), (Winston) Wai-Chi Fang, (National Chiao Tung University, Taiwan), Kun Chang Lee (Sungkyunkwan Univ., South Korea), Andy Marsh (HoIP, ICET), Layne T. Watson (Virginia Polytechnic Institute & State Univ., Virginia, USA), and Lotfi A. Zadeh (Univ. of California, Berkeley, USA).

Most importantly, we wish to express again our sincere gratitude and respect towards Professor Hamid R. Arabnia (Univ. of Georgia, USA), General Chair of all WORLDCOMP conferences, for his excellent and tireless support, organization and coordination of all affiliated events. His exemplary and professional effort in 2011 and all the years before in the WORLDCOMP steering committee makes these events possible!

Thank you all for your contribution to DMIN'11! We hope that you will experience a stimulating conference with many opportunities for future contacts, research and applications.

Robert Stahlbock  
DMIN'11 General Conference Chair



# Contents

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

<b>Three Different Paradigms for Interactive Data Clustering</b>	<b>3</b>
<i>Terje Kristensen, Vemund Jakobsen</i>	
<b>Noise-Tolerant Active Learning</b>	<b>10</b>
<i>Tsutomu Osoda, Satoru Miyano</i>	
<b>Breast Cancer Risk Score A Data Mining Approach to Improve Readability</b>	<b>15</b>
<i>Emilien Gauthier, Laurent Brisson, Philippe Lenca, Stephane Ragusa</i>	
<b>Soft-sensors for Real-time Monitoring and Control of a Black Liquor Concentration Process</b>	<b>22</b>
<i>Mouloud Amazouz, radu Platon</i>	
<b>Ant Colony Optimization with Ants' Individual Memories</b>	<b>28</b>
<i>Hiroki Inoue, Yasuhiko Kato</i>	
<b>A Prediction Model for Recognition of Bad Credit Customers in Saman Bank Using Neural Networks</b>	<b>34</b>
<i>Masoud Yaghini, Toktam Zhiyan, Mehdi Fallahi</i>	
<b>A New Approach for Handling Numeric Ranges for Graph-Based Knowledge Discovery</b>	<b>40</b>
<i>Oscar Romero , Lawrence Holder, Jesus Gonzalez</i>	
<b>A Framework for Detecting Vulnerable, Cascaded Fuzzy Cycles in the Carbon Chain</b>	<b>47</b>
<i>James Buckley, Jennifer Seitzer</i>	
<b>Reliability Analysis of Markov Blanket Learning Algorithms (1996-2010)</b>	<b>53</b>
<i>Shunkai Fu, Michel Desmarais, Weibin Chen</i>	
<b>Sobek: a Text Mining Tool for Educational Applications</b>	<b>59</b>
<i>Eliseo Reategui, Miriam Klemann, Daniel Epstein, Alexandre Lorenzatti</i>	
<b>A Novel Soft Computing Hybrid for Data Imputation</b>	<b>65</b>
<i>Ankaiah Narravula, Ravi Vadlamani</i>	
<b>Bankruptcy Prediction in Banks by Principal Component Analysis Threshold Accepting Trained Wavelet Neural Network Hybrid</b>	<b>71</b>
<i>Vasu Madireddi, Ravi Vadlamani</i>	

<b>A Real Application on Non-technical Losses Detection: The MIDAS Project</b>	<b>77</b>
<i>Juan I. Guerrero, Carlos Leon, Felix Biscarri, Inigo Monedero, Jesus Biscarri, Rocio Millan</i>	
<b>Capstone Project: Event Monitoring and Alerting System</b>	<b>84</b>
<i>Wook-Sung Yoo, Geetha Rajgopalan</i>	
<b>Constrained Nonnegative Matrix Factorization for Data Privacy</b>	<b>88</b>
<i>Nirmal Thapa, Lian Liu, Pengpeng Lin, Jie Wang, Jun Zhang</i>	
<b>An Optimization Framework for Process Discovery Algorithms</b>	<b>94</b>
<i>Ton Weijters</i>	
<b>Facial Nerve Stream Trajectory Data Modelling and Visualization</b>	<b>101</b>
<i>Jalel Akaichi, Hanen Bouali, Zeineb Dhouioui</i>	
<b>Domain Specific Services for Continuous Diagnoses in the Context of Ambient Assisted Living AAL</b>	<b>106</b>
<i>Bjoern-Helge Busch, Ralph Welge</i>	
<b>Position of Gateway Drugs in the Spectrum of Adolescent Drug-Use Initiation in Indiana - Relevance of Market Basket Analysis of Data Mining in Detection of Common Substance-Use Initiation Sequences</b>	<b>113</b>
<i>Ahmed YoussefAgha, Wasantha Jayawardene</i>	
<b>SESSION: SEGMENTATION, CLUSTERING, ASSOCIATION</b>	
<b>Clustering Approach Based On Von Neumann Topology Artificial Bee Colony Algorithm</b>	<b>121</b>
<i>Wenping Zou, Yunlong Zhu, Hanning Chen, Tao Ku</i>	
<b>Hierarchical Random Graphs for Networks with Weighted Edges and Multiple Edge Attributes</b>	<b>128</b>
<i>David Allen, Tsai-Ching Lu, Dave Huber, Hankyu Moon</i>	
<b>A Two-Stage Algorithm for Data Clustering</b>	<b>135</b>
<i>Abdolreza Hatamlou, Salwani Abdullah</i>	
<b>A Binary Based Approach for Generating Association Rules</b>	<b>140</b>
<i>Mohamed El Hadi Benelhadj, Khediya Arour, Mahmoud Boufaida, Yahya Slimani</i>	
<b>Casino Fraud Data Mining</b>	<b>147</b>
<i>Robert Woodley, Warren Noll, Kevin Shallenberger</i>	



<b>A Case Study on Clustering and Mining Business Processes from a University</b>	<b>153</b>
<i>Pedro Esposito, Marco Vaz, Jano Souza, Luciano Terres</i>	
<b>Centrality Preservation in Anonymized Social Networks</b>	<b>160</b>
<i>Traian Marius Truta, Alina Campan, Ashley Gasmi, Nicholas Cooper, Andrew Elstun</i>	
<b>Design of Customer Behavior Analysis Model in Automobile Marketing</b>	<b>167</b>
<i>Yuanyuan Mao, Lan Huang, Guishen Wang, Shuxue Zou</i>	
<b>An Approach to Selecting Proper Dimensions for Noisy Data</b>	<b>172</b>
<i>Yong Shi, Jerry Meisner</i>	
<b>Adaptive Neuro Fuzzy Networks based on Quantum Subtractive Clustering</b>	<b>176</b>
<i>Ali Mousavi, Mehrdad Jalali, Mahdi Yaghoubi</i>	
<b>A Clustering Approach to Unsupervised Attack Detection in Collaborative Recommender Systems</b>	<b>181</b>
<i>Runa Bhaumik, Bamshad Mobasher, Robin Burke</i>	
<b>Stable Clustering of Temporal Gene Expression Data</b>	<b>188</b>
<i>Gaolin Zheng, Guowang Mu, Chung-Hao Chen, Xinyu Huang</i>	
<b>Heuristic Approaches for Embedded Processor System Size Reduction</b>	<b>192</b>
<i>Rahul Dixit, Harpreet Singh</i>	
<b>Probabilistic Vector Machines</b>	<b>198</b>
<i>Henri Luchian, Andrei Sucila</i>	
<b>Mining Association Rules from Responded Questionnaire of Sanitary Education Guidance</b>	<b>203</b>
<i>Yo-Ping Huang, Zheng-Hong Deng, Shan-Shan Wang</i>	
<b>A New Term Weighting Scheme For Document Clustering</b>	<b>210</b>
<i>Keerthiram Murugesan, Jun Zhang</i>	
<b>A New Approach to Present Prototypes in Clustering of Time Series</b>	<b>214</b>
<i>Saeed R. Aghabozorgi, Teh Ying Wah, Amineh Amini, Mahmud R. Saybani</i>	
<b>SESSION: REGRESSION, CLASSIFICATION</b>	
<b>Threshold Value Based Traffic Congestion Identification Method</b>	<b>223</b>
<i>Zhanquan Sun, Weidong Gu, Jinqiao Feng, Xiaomin Zhu</i>	
<b>Comparison of Single Image Processing and Bilateral Image Feature Subtraction in Breast Cancer Detection</b>	<b>230</b>
<i>Aijuan Dong, Sinanovic Senad</i>	

<b>Simple R-Tree for Temporal Searches</b>	<b>236</b>
<i>Paul Te Braak, Richi Nayak</i>	
<b>On Sample Selection Bias in Large-Scale Online Stream Mining: a Model Indexing Approach</b>	<b>240</b>
<i>Xiong Deng, Moustafa Ghanem, Yike Guo</i>	
<b>HIOPGA: A New Hybrid Metaheuristic Algorithm to Train Feedforward Neural Networks for Prediction</b>	<b>248</b>
<i>Masoud Yaghini, Mohammad M. Khoshraftar, Mehdi Fallahi</i>	
<b>Incremental Classification Based on Association Rules Algorithm (ICBA)</b>	<b>255</b>
<i>Sararak Tanarat, Worapoj Kreesuradej</i>	
<b>Constrained Multi-Label Classification: A Semidefinite Programming Approach</b>	<b>259</b>
<i>Hui Wu, Guangzhi Qu, Hui Zhang, Craig Hartrick</i>	
<b>An Empirical Study of Class Noise Impacts on Supervised Learning Algorithms and Measures</b>	<b>266</b>
<i>Victor Sheng, Rahul Tada, Abhinav Atla</i>	
<b>A Novel Protein Secondary Structure Intelligent Prediction System</b>	<b>273</b>
<i>Bingru Yang</i>	
<b>Bankruptcy Prediction with Missing Data</b>	<b>279</b>
<i>Qi Yu, Yoan Miche, Amaury Lendasse, Eric Severin</i>	
<b>An EM-based Multi-Step Piecewise Surface Regression Learning Algorithm</b>	<b>286</b>
<i>Juan Luo, Alexander Brodsky</i>	
<b>SESSION: EXPLORATIVE DATA MINING, DATA PREPROCESSING, FEATURE SELECTION</b>	
<b>A Computerized Feature Reduction Using Principal Component Analysis for Accident Duration Forecasting on Freeway</b>	<b>295</b>
<i>Ying Lee</i>	
<b>Example Labeling Difficulty within Repeated Labeling</b>	<b>301</b>
<i>Victor Sheng</i>	
<b>Mining Frequent Item Sets Efficiently by Using Compression Techniques</b>	<b>308</b>
<i>Selim Mimaroglu, Cagri Cubukcu, Emin Aksehirli, Ertunc Erdil</i>	

**Modeling Functional Outliers for High Frequency Time Series Forecasting with Neural Networks: An Empirical Evaluation for Electricity Load Data** 313  
*Nikolaos Kourentzes*

**Feature Selection with Hybrid Mutual Information and Genetic Algorithm** 320  
*Vahid Chahkandi, Mehrdad Jalali, Mahsa Mirshahi, Ali Hosseini*

**Finding Perfect-Predictor Feature Sets for Supervised Classification Using Genetic Algorithms** 325  
*Alexander Liu, Cheryl Martin*

**A Novel Knowledge-discovering Approach from Massive Data** 332  
*Heni Bouhamed, Ahmed Rebai, Thierry Lecroq, Maher Jaoua*

**Improved Interpretability of the Unified Distance Matrix with Connected Components** 338  
*Lutz Hamel, Chris W. Brown*

**On Selecting the Number of Bins for a Histogram** 344  
*Sai Venu Gopal Lolla, Lawrence L. Hoberock*

#### **SESSION: WEB AND TEXT MINING**

**A Secure Knowledge Discovery Framework for Clinical Informatics** 353  
*Yueh-Hsun Shih, Chung-Yueh Lien, Chi-Hsien Chen, Chia-Hung Hsiao, Woei-Chyn Chu*

**Pattern-based Aggregation of Named Entity Extractors** 357  
*Tracy Lemmond, Paul Kidwell, Kofi Boakye, Nathan Perry, Joseph Guensche, John Nitao, William Hanley, Ryan Prenger, Ron Glaser*

**Sentiment Detection with Character n-Grams** 364  
*Tino Hartmann, Sebastian Klenk, Andre Burkovski, Gunther Heidemann*

**Objective Words Can Improve Sentiment Classification for Word of Mouth** 369  
*Chihli Hung, Hao-Kai Lin, Chih-Fong Tsai*

**Privacy-Preserving Profiling** 372  
*Thomas Barnard, Adam Prugel-Bennett*

#### **SESSION: FEATURE SELECTION + CLUSTERING METHODS + TESTING APPLICATIONS**

**Perspective of Feature Selection Techniques in Bioinformatics** 381  
*Satish Kumar, Mohammad Khalid Siddiqui*

<b>Modularity and Spectral Co-Clustering for Categorical Data</b>	<b>386</b>
<i>Lazhar Labiod, Mohamed Nadif</i>	
<b>Statistical Procedure For Simultaneous Testing of Many Hypothesis</b>	<b>393</b>
<i>Gurpreet Bawa</i>	