

Honghua Dai Ramakrishnan Srikant
Chengqi Zhang (Eds.)

Advances in Knowledge Discovery and Data Mining

8th Pacific-Asia Conference, PAKDD 2004
Sydney, Australia, May 26-28, 2004
Proceedings



Springer

Table of Contents

Invited Speeches

Mining of Evolving Data Streams with Privacy Preservation	1
<i>Philip S. Yu</i>	

Data Mining Grand Challenges	2
<i>Usama Fayyad</i>	

Session 1A: Classification (I)

Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms	3
<i>Remco R. Bouckaert, Eibe Frank</i>	

Spectral Energy Minimization for Semi-supervised Learning	13
<i>Chun-hung Li, Zhi-li Wu</i>	

Discriminative Methods for Multi-labeled Classification	22
<i>Shantanu Godbole, Sunita Sarawagi</i>	

Session 1B: Clustering (I)

Subspace Clustering of High Dimensional Spatial Data with Noises	31
<i>Chih-Ming Hsu, Ming-Syan Chen</i>	

Constraint-Based Graph Clustering through Node Sequencing and Partitioning	41
<i>Yu Qian, Kang Zhang, Wei Lai</i>	

Mining Expressive Process Models by Clustering Workflow Traces	52
<i>Gianluigi Greco, Antonella Guzzo, Luigi Pontieri, Domenico Saccà</i>	

Session 1C: Association Rules (I)

CMTreMiner: Mining Both Closed and Maximal Frequent Subtrees	63
<i>Yun Chi, Yirong Yang, Yi Xia, Richard R. Muntz</i>	

Secure Association Rule Sharing	74
<i>Stanley R.M. Oliveira, Osmar R. Zaiane, Yücel Saygin</i>	

Self-Similar Mining of Time Association Rules	86
<i>Daniel Barbará, Ping Chen, Zohreh Nazeri</i>	

Session 2A: Novel Algorithms (I)

ParaDualMiner: An Efficient Parallel Implementation
of the DualMiner Algorithm 96
Roger M.H. Ting, James Bailey, Kotagiri Ramamohanarao

A Novel Distributed Collaborative Filtering Algorithm
and Its Implementation on P2P Overlay Network 106
Peng Han, Bo Xie, Fan Yang, Jiajun Wang, Ruimin Shen

An Efficient Algorithm for Dense Regions Discovery
from Large-Scale Data Streams 116
Andy M. Yip, Edmond H. Wu, Michael K. Ng, Tony F. Chan

Blind Data Linkage Using n -gram Similarity Comparisons..... 121
Tim Churches, Peter Christen

Condensed Representation of Emerging Patterns 127
Arnaud Soulet, Bruno Crémilleux, François Rioult

Session 2B: Association (II)

Discovery of Maximally Frequent Tag Tree Patterns
with Contractible Variables from Semistructured Documents 133
*Tetsuhiro Miyahara, Yusuke Suzuki, Takayoshi Shoudai,
Tomoyuki Uchida, Kenichi Takahashi, Hiroaki Ueda*

Mining Term Association Rules for Heuristic Query Construction 145
Zhenxing Qin, Li Liu, Shichao Zhang

FP-Bonsai: The Art of Growing and Pruning Small FP-Trees 155
Francesco Bonchi, Bart Goethals

Mining Negative Rules Using GRD 161
Dhananjay R. Thiruvady, Geoff I. Webb

Applying Association Rules for Interesting Recommendations
Using Rule Templates 166
Jiye Li, Bin Tang, Nick Cercone

Session 2C: Classification (II)

Feature Extraction and Classification System for Nonlinear
and Online Data 171
Byung Joo Kim, Il Kon Kim, Kwang Baek Kim

A Metric Approach to Building Decision Trees
Based on Goodman-Kruskal Association Index 181
Dan A. Simovici, Szymon Jaroszewicz

DRC-BK: Mining Classification Rules with Help of SVM	191
<i>Yang Zhang, Zhanhuai Li, Yan Tang, Kebin Cui</i>	
A New Data Mining Method Using Organizational Coevolutionary Mechanism	196
<i>Jing Liu, Weicai Zhong, Fang Liu, Licheng Jiao</i>	
Noise Tolerant Classification by Chi Emerging Patterns	201
<i>Hongjian Fan, Kotagiri Ramamohanarao</i>	
The Application of Emerging Patterns for Improving the Quality of Rare-Class Classification	207
<i>Hamad Alhammad, Kotagiri Ramamohanarao</i>	
Session 3A: Event Mining, Anomaly Detection, and Intrusion Detection	
Finding Negative Event-Oriented Patterns in Long Temporal Sequences	212
<i>Xingzhi Sun, Maria E. Orlowska, Xue Li</i>	
OBE: Outlier by Example	222
<i>Cui Zhu, Hiroyuki Kitagawa, Spiros Papadimitriou, Christos Faloutsos</i>	
Temporal Sequence Associations for Rare Events	235
<i>Jie Chen, Hongxing He, Graham Williams, Huidong Jin</i>	
Summarization of Spacecraft Telemetry Data by Extracting Significant Temporal Patterns	240
<i>Takehisa Yairi, Shiro Ogasawara, Koichi Hori, Shinichi Nakasuka, Naoki Ishihama</i>	
An Extended Negative Selection Algorithm for Anomaly Detection	245
<i>Xiaoshu Hang, Honghua Dai</i>	
Adaptive Clustering for Network Intrusion Detection	255
<i>Joshua Oldmeadow, Siddarth Ravinutala, Christopher Leckie</i>	
Session 3B: Ensemble Learning	
Ensembling MML Causal Discovery	260
<i>Honghua Dai, Gang Li, Zhi-Hua Zhou</i>	
Logistic Regression and Boosting for Labeled Bags of Instances	272
<i>Xin Xu, Eibe Frank</i>	
Fast and Light Boosting for Adaptive Mining of Data Streams	282
<i>Fang Chu, Carlo Zaniolo</i>	

Compact Dual Ensembles for Active Learning	293
<i>Amit Mandvikar, Huan Liu, Hiroshi Motoda</i>	
On the Size of Training Set and the Benefit from Ensemble.....	298
<i>Zhi-Hua Zhou, Dan Wei, Gang Li, Honghua Dai</i>	
Session 3C: Bayesian Network and Graph Mining	
Identifying Markov Blankets Using Lasso Estimation	308
<i>Gang Li, Honghua Dai, Yiqing Tu</i>	
Selective Augmented Bayesian Network Classifiers Based on Rough Set Theory	319
<i>Zhihai Wang, Geoffrey I. Webb, Fei Zheng</i>	
Using Self-Consistent Naive-Bayes to Detect Masquerades	329
<i>Kwong H. Yung</i>	
DB-Subdue: Database Approach to Graph Mining	341
<i>Sharma Chakravarthy, Ramji Beera, Ramanathan Balachandran</i>	
Session 3D: Text Mining (I)	
Finding Frequent Structural Features among Words in Tree-Structured Documents	351
<i>Tomoyuki Uchida, Tomonori Mogawa, Yasuaki Nakamura</i>	
Exploring Potential of Leave-One-Out Estimator for Calibration of SVM in Text Mining	361
<i>Adam Kowalczyk, Bhavani Raskutti, Herman Ferrá</i>	
Classifying Text Streams in the Presence of Concept Drifts	373
<i>Gabriel Pui Cheong Fung, Jeffrey Xu Yu, Hongjun Lu</i>	
Using Cluster-Based Sampling to Select Initial Training Set for Active Learning in Text Classification	384
<i>Jaeho Kang, Kwang Ryel Ryu, Hyuk-Chul Kwon</i>	
Spectral Analysis of Text Collection for Similarity-Based Clustering	389
<i>Wenyuan Li, Wee-Keong Ng, Ee-Peng Lim</i>	
Session 4A: Clustering (II)	
Clustering Multi-represented Objects with Noise	394
<i>Karin Kailing, Hans-Peter Kriegel, Alexey Pryakhin, Matthias Schubert</i>	
Providing Diversity in K-Nearest Neighbor Query Results	404
<i>Anoop Jain, Parag Sarda, Jayant R. Haritsa</i>	

Cluster Structure of K -means Clustering via Principal Component Analysis	414
<i>Chris Ding, Xiaofeng He</i>	
Combining Clustering with Moving Sequential Pattern Mining: A Novel and Efficient Technique	419
<i>Shuai Ma, Shiwei Tang, Dongqing Yang, Tengjiao Wang, Jinjiang Han</i>	
An Alternative Methodology for Mining Seasonal Pattern Using Self-Organizing Map	424
<i>Denny Lee, Vincent C.S. Lee</i>	
Session 4B: Association (III)	
ISM: Item Selection for Marketing with Cross-Selling Considerations	431
<i>Raymond Chi-Wing Wong, Ada Wai-Chee Fu</i>	
Efficient Pattern-Growth Methods for Frequent Tree Pattern Mining	441
<i>Chen Wang, Mingsheng Hong, Jian Pei, Haofeng Zhou, Wei Wang, Baile Shi</i>	
Mining Association Rules from Structural Deltas of Historical XML Documents	452
<i>Ling Chen, Sourav S. Bhowmick, Liang-Tien Chia</i>	
Data Mining Proxy: Serving Large Number of Users for Efficient Frequent Itemset Mining	458
<i>Zhiheng Li, Jeffrey Xu Yu, Hongjun Lu, Yabo Xu, Guimei Liu</i>	
Session 4C: Novel Algorithms (II)	
Formal Approach and Automated Tool for Translating ER Schemata into OWL Ontologies	464
<i>Zhuoming Xu, Xiao Cao, Yisheng Dong, Wenping Su</i>	
Separating Structure from Interestingness	476
<i>Taneli Mielikäinen</i>	
Exploiting Recurring Usage Patterns to Enhance Filesystem and Memory Subsystem Performance	486
<i>Benjamin Rutt, Srinivasan Parthasarathy</i>	
Session 4D: Multimedia Mining	
Automatic Text Extraction for Content-Based Image Indexing	497
<i>Keechul Jung, Eun Yi Kim</i>	
Peculiarity Oriented Analysis in Multi-people Tracking Images	508
<i>Muneaki Ohshima, Ning Zhong, Y. Y. Yao, Shinichi Murata</i>	

AutoSplit: Fast and Scalable Discovery of Hidden Variables in Stream and Multimedia Databases 519
Jia-Yu Pan, Hiroyuki Kitagawa, Christos Faloutsos, Masafumi Hamamoto

Session 5A: Text Mining and Web Mining (II)

Semantic Sequence Kin: A Method of Document Copy Detection 529
Jun-Peng Bao, Jun-Yi Shen, Xiao-Dong Liu, Hai-Yan Liu, Xiao-Di Zhang

Extracting Citation Metadata from Online Publication Lists Using BLAST 539
I-Ane Huang, Jan-Ming Ho, Hung-Yu Kao, Wen-Chang Lin

Mining of Web-Page Visiting Patterns with Continuous-Time Markov Models 549
Qiming Huang, Qiang Yang, Joshua Zhexue Huang, Michael K. Ng

Discovering Ordered Tree Patterns from XML Queries 559
Yi Chen

Predicting Web Requests Efficiently Using a Probability Model 564
Shanchan Wu, Wenyuan Wang

Session 5B: Statistical Methods, Sequential Data Mining, and Time Series Mining

CCMine: Efficient Mining of Confidence-Closed Correlated Patterns 569
Won-Young Kim, Young-Koo Lee, Jiawei Han

A Conditional Probability Distribution-Based Dissimilarity Measure for Categorical Data 580
Le Si Quang, Ho Tu Bao

Learning Hidden Markov Model Topology Based on KL Divergence for Information Extraction 590
Kwok-Chung Au, Kwok-Wai Cheung

A Non-parametric Wavelet Feature Extractor for Time Series Classification 595
Hui Zhang, Tu Bao Ho, Mao Song Lin

Rules Discovery from Cross-Sectional Short-Length Time Series 604
Kedong Luo, Jianmin Wang, Jianguang Sun

Session 5C: Novel Algorithms (III)

Constraint-Based Mining of Formal Concepts in Transactional Data	615
<i>Jérémy Besson, Céline Robardet, Jean-François Boulicaut</i>	
Towards Optimizing Conjunctive Inductive Queries	625
<i>Johannes Fischer, Luc De Raedt</i>	
Febri – A Parallel Open Source Data Linkage System	638
<i>Peter Christen, Tim Churches, Markus Hegland</i>	
A General Coding Method for Error-Correcting Output Codes	648
<i>Yan-huang Jiang, Qiang-li Zhao, Xue-jun Yang</i>	
Discovering Partial Periodic Patterns in Discrete Data Sequences	653
<i>Huiping Cao, David W. Cheung, Nikos Mamoulis</i>	

Session 5D: Biomedical Mining

Conceptual Mining of Large Administrative Health Data	659
<i>Tatiana Semenova, Markus Hegland, Warwick Graco, Graham Williams</i>	
A Semi-automatic System for Tagging Specialized Corpora	670
<i>Ahmed Amrani, Yves Kodratoff, Oriane Matte-Tailliez</i>	
A Tree-Based Approach to the Discovery of Diagnostic Biomarkers for Ovarian Cancer	682
<i>Jinyan Li, Kotagiri Ramamohanarao</i>	
A Novel Parameter-Less Clustering Method for Mining Gene Expression Data	692
<i>Vincent Shin-Mu Tseng, Ching-Pin Kao</i>	
Extracting and Explaining Biological Knowledge in Microarray Data	699
<i>Paul J. Kennedy, Simeon J. Simoff, David Skillicorn, Daniel Catchpoole</i>	
Further Applications of a Particle Visualization Framework	704
<i>Ke Yin, Ian Davidson</i>	
Author Index	711