Advances in Knowledge Discovery and Data Mining

7th Pacific-Asia Conference, PAKDD 2003
Seoul, Korea, April 30 – May 2, 2003
Proceedings
Table of Contents

Industrial Papers (Invited)

Data Mining as an Automated Service ........................................ 1
    P.S. Bradley

Trends and Challenges in the Industrial Applications of KDD .......... 14
    Ramasamy Uthurusamy

Session 1A: Stream Mining I

Finding Event-Oriented Patterns in Long Temporal Sequences .......... 15
    Xingzhi Sun, Maria E. Orlowska, Xiaofang Zhou

Mining Frequent Episodes for Relating Financial Events and Stock Trends ........................................ 27
    Anny Ng, Ada Wai-chee Fu

Session 1B: Graph Mining

An Efficient Algorithm of Frequent Connected Subgraph Extraction ........................................ 40
    Mingsheng Hong, Hao Feng Zhou, Wei Wang, Baile Shi

Classifier Construction by Graph-Based Induction for Graph-Structured Data ........................................ 52
    Warodom Geamsakul, Takashi Matsuda, Tetsuya Yoshida,
    Hiroshi Motoda, Takashi Washio

Session 1C: Clustering I

Comparison of the Performance of Center-Based Clustering Algorithms ........................................ 63
    Bin Zhang

Automatic Extraction of Clusters from Hierarchical Clustering Representations ........................................ 75
    Jörg Sander, Xuejie Qin, Zhiyong Lu, Nan Niu, Alex Kovarsky

Session 2A: Text Mining

Large Scale Unstructured Document Classification Using Unlabeled Data and Syntactic Information ........................................ 88
    Seong-Bae Park, Byoung-Tak Zhang
## Session 2B: Bio Mining

Mining Patterns of Dyspepsia Symptoms Across Time Points Using Constraint Association Rules

_Annie Lau, Siew Siew Ong, Ashesh Mahidadia, Achim Hoffmann, Johanna Westbrook, Tatjana Zrimec_

Predicting Protein Structural Class from Closed Protein Sequences

_N. Rattanakronkul, T. Wattarujeekrit, K. Waiyamai_

Learning Rules to Extract Protein Interactions from Biomedical Text

_Tu Minh Phuong, Doheon Lee, Kwang Hyung Lee_

Predicting Protein Interactions in Human by Homologous Interactions in Yeast

_Hyongguen Kim, Jong Park, Kyungsook Han_

## Session 2C: Web Mining

Mining the Customer's Up-To-Moment Preferences for E-commerce Recommendation

_Yi-Dong Shen, Qiang Yang, Zhong Zhang, Hongjun Lu_

A Graph-Based Optimization Algorithm for Website Topology Using Interesting Association Rules

_Edmond H. Wu, Michael K. Ng_

A Markovian Approach for Web User Profiling and Clustering

_Younes Hafri, Chabane Djeraba, Peter Stanchev, Bruno Bachimont_

Extracting User Interests from Bookmarks on the Web

_Jason J. Jung, Geun-Sik Jo_

## Session 3A: Stream Mining II

Mining Frequent Instances on Workflows

_Gianluigi Greco, Antonella Guzzo, Giuseppe Manco, Domenico Saccà_
Real Time Video Data Mining for Surveillance Video Streams .......... 222
   JungHwan Oh, JeongKyu Lee, Sanjaykumar Kote

Distinguishing Causal and Acausal Temporal Relations ................. 234
   Kamran Karimi, Howard J. Hamilton

Session 3B: Bayesian Networks

Online Bayes Point Machines .......................................... 241
   Edward Harrington, Ralf Herbrich, Jyrki Kivinen, John Platt,
   Robert C. Williamson

Exploiting Hierarchical Domain Values for Bayesian Learning .......... 253
   Yiqiu Han, Wai Lam

A New Restricted Bayesian Network Classifier ........................ 265
   Hongbo Shi, Zhihai Wang, Geoff Webb, Houkuan Huang

Session 3C: Clustering II

AGRID: An Efficient Algorithm for Clustering Large
High-Dimensional Datasets ............................................ 271
   Zhao Yanchang, Song Junde

Multi-level Clustering and Reasoning about Its Clusters Using
Region Connection Calculus ........................................... 283
   Ickjai Lee, Mary-Anne Williams

An Efficient Cell-Based Clustering Method for Handling Large,
High-Dimensional Data ................................................ 295
   Jae-Woo Chang

Session 4A: Association Rules I

Enhancing SWF for Incremental Association Mining
by Itemset Maintenance .................................................. 301
   Chia-Hui Chang, Shi-Hsan Yang

Reducing Rule Covers with Deterministic Error Bounds ................ 313
   Vikram Pudi, Jayant R. Haritsa

Evolutionary Approach for Mining Association Rules on
Dynamic Databases ....................................................... 325
   P. Deepa Shenoy, K.G. Srinivasa, K.R. Venugopal, L.M. Patnaik
Session 4B: Semi-structured Data Mining

Position Coded Pre-order Linked WAP-Tree for Web Log Sequential Pattern Mining ........................................... 337
Yi Lu, C.I. Ezeife

An Integrated System of Mining HTML Texts and Filtering Structured Documents ........................................... 350
Bo-Hyun Yun, Myung-Eun Lim, Soo-Hyun Park

A New Sequential Mining Approach to XML Document Similarity Computation ........................................... 356
Ho-pong Leung, Fu-lai Chung, Stephen Chi-fai Chan

Session 4C: Classification I

Optimization of Fuzzy Rules for Classification Using Genetic Algorithm ......................................................... 363
Myung Won Kim, Joung Woo Ryu, Samkeun Kim, Joong Geun Lee

Fast Pattern Selection for Support Vector Classifiers ......................................................................................... 376
Hyunjung Shin, Sungzoon Cho

Averaged Boosting: A Noise-Robust Ensemble Method .......................................................... 388
Yongdai Kim

Improving Performance of Decision Tree Algorithms with Multi-edited Nearest Neighbor Rule .................. 394
Chen-zhou Ye, Jie Yang, Li-xiu Yao, Nian-yi Chen

Session 5A: Data Analysis

HOT: Hypergraph-Based Outlier Test for Categorical Data ............................................................. 399
Li Wei, Weining Qian, Aoying Zhou, Wen Jin, Jeffrey X. Yu

A Method for Aggregating Partitions, Applications in K.D.D. .......................................................... 411
Pierre-Emmanuel Jouve, Nicolas Nicoloyannis

Efficiently Computing Iceberg Cubes with Complex Constraints through Bounding ........................ 423
Pauline L.H. Chou, Xiuzhen Zhang

Extraction of Tag Tree Patterns with Contractible Variables from Irregular Semistructured Data ........... 430
Tetsuhiro Miyahara, Yusuke Suzuki, Takayoshi Shoudai, Tomoyuki Uchida, Sachio Hirokawa, Kenichi Takahashi, Hiroaki Ueda
Session 5B: Association Rules II

Step-By-Step Regression: A More Efficient Alternative for Polynomial Multiple Linear Regression in Stream Cube .......... 437
   Chao Liu, Ming Zhang, Minrui Zheng, Yixin Chen

Progressive Weighted Miner: An Efficient Method for Time-Constraint Mining .................................................. 449
   Chang-Hung Lee, Jian Chih Ou, Ming-Syan Chen

Mining Open Source Software (OSS) Data Using Association Rules Network .......................................................... 461
   Sanjay Chawla, Bavani Arunasalam, Joseph Davis

Parallel FP-Growth on PC Cluster ............................................. 467
   Iko Pramudiono, Masaru Kitsuregawa

Session 5C: Feature Selection

Active Feature Selection Using Classes .................................... 474
   Huan Liu, Lei Yu, Manoranjan Dash, Hiroshi Motoda

Electricity Based External Similarity of Categorical Attributes ........ 486
   Christopher R. Palmer, Christos Faloutsos

Weighted Proportional k-Interval Discretization for Naive-Bayes Classifiers ...................................................... 501
   Ying Yang, Geoffrey I. Webb

Dealing with Relative Similarity in Clustering:
   An Indiscernibility Based Approach ...................................... 513
   Shoji Hirano, Shusaku Tsumoto

Session 6A: Stream Mining III

Considering Correlation between Variables to Improve Spatiotemporal Forecasting ........................................ 519
   Zhigang Li, Liangang Liu, Margaret H. Dunham

Correlation Analysis of Spatial Time Series Datasets: A Filter-and-Refine Approach ................................... 532
   Pusheng Zhang, Yan Huang, Shashi Shekhar, Vipin Kumar

When to Update the Sequential Patterns of Stream Data? ............. 545
   Qingguo Zheng, Ke Xu, Shilong Ma

Session 6B: Clustering III

A New Clustering Algorithm for Transaction Data via Caucus .......... 551
   Jinmei Xu, Hui Xiong, Sam Yuan Sung, Vipin Kumar
<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBRS: A Density-Based Spatial Clustering Method with Random Sampling</td>
<td>563</td>
</tr>
<tr>
<td>Xin Wang, Howard J. Hamilton</td>
<td></td>
</tr>
<tr>
<td>Optimized Clustering for Anomaly Intrusion Detection</td>
<td>576</td>
</tr>
<tr>
<td>Sang Hyun Oh, Won Suk Lee</td>
<td></td>
</tr>
<tr>
<td><strong>Session 6C: Classification II</strong></td>
<td></td>
</tr>
<tr>
<td>Finding Frequent Subgraphs from Graph Structured Data with</td>
<td>582</td>
</tr>
<tr>
<td>Geometric Information and Its Application to Lossless Compression</td>
<td></td>
</tr>
<tr>
<td>Yuko Itokawa, Tomoyuki Uchida, Takayoshi Shoudai,</td>
<td></td>
</tr>
<tr>
<td>Tetsuhiro Miyahara, Yasuaki Nakamura</td>
<td></td>
</tr>
<tr>
<td>Upgrading ILP Rules to First-Order Bayesian Networks</td>
<td>595</td>
</tr>
<tr>
<td>Ratthachat Chatpatanasiri, Boonserm Kijsirikul</td>
<td></td>
</tr>
<tr>
<td>A Clustering Validity Assessment Index</td>
<td>602</td>
</tr>
<tr>
<td>Youngok Kim, Soowon Lee</td>
<td></td>
</tr>
<tr>
<td><strong>Author Index</strong></td>
<td>609</td>
</tr>
</tbody>
</table>