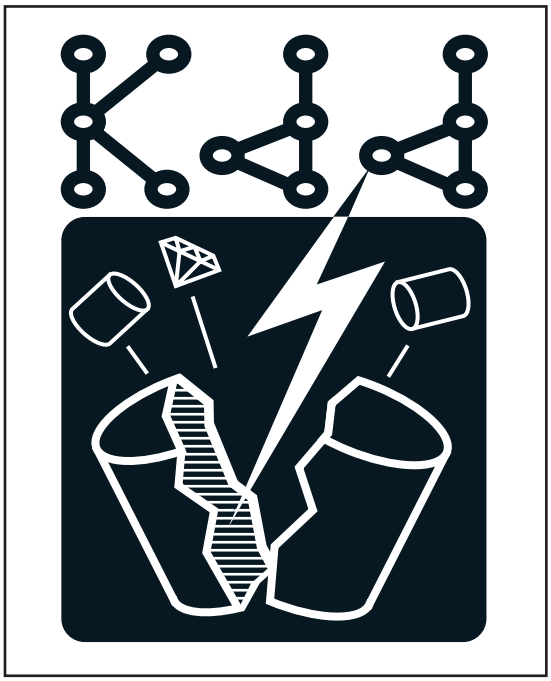


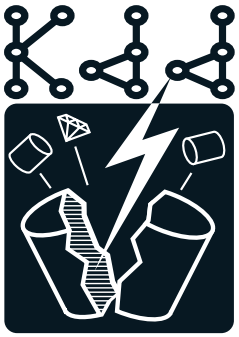
The Fourth International Conference on Knowledge Discovery and Data Mining

August 27–31 1998, New York, New York

Sponsored by the American Association for Artificial Intelligence

Cosponsored by
Epiphany
Intel Corporation
Knowledge Stream Partners
Magnify, Inc.
Microsoft Corporation
National Institute of Statistical Sciences
SAS Institute Inc.
Silicon Graphics, Inc.

In cooperation with the
Twenty-Fourth Annual International Conference on
Very Large Databases (VLDB '98)



Welcome to KDD-98!

KDD-98 Organization

General Conference Chair
Gregory Piatetsky-Shapiro, *Knowledge Stream Partners*

Program Cochairs
Rakesh Agrawal, *IBM Almaden Research Center*
Paul Stolorz, *Jet Propulsion Laboratory*

Publicity Chair
Foster Provost, *Bell Atlantic Science and Technology*

Tutorial Chair
Padhraic Smyth, *University of California, Irvine*

Panel Chair
Willi Kloesgen, *GMD, Germany*

Workshops Chair
Ronny Kohavi, *Silicon Graphics*

Exhibits Chair
Ismail Parsa, *Epsilon*

Poster Sessions Chair
David Jensen, U*niversity of Massachusetts, Amherst*

Local Arrangements Chair
Kyusoek Shim, *Bell Laboratories*

Sponsorship Chair
Ramasamy Uthurusamy, *General Motors Corporation*

Webmaster
Alexander Gray

Program Committee Members
Rakesh Agrawal, *IBM Almaden Research Center*
Tej Anand, *Golden Books Family Entertainment*
Chid Apte, *IBM TJ Watson Research Center*
Andreas Arning, *IBM, Germany*
Roberto Bayardo, *IBM Almaden Research Center*
Carla Brodley, *Purdue University*
Wray Buntine, *Ultimode Systems*
Michael Burl, *Jet Propulsion Laboratory*
Soumen Chakrabarti, *IBM Almaden Research Center*
Ernest Chan, *Credit Suisse First Boston*
Surajit Chaudhuri, *Microsoft Research*
Corinna Cortes, *AT&T Laboratories*
Bruce Croft, *University of Massachusetts, Amherst*
Umeshwar Dayal, *Hewlett Packard Laboratories*
Pedro Domingos, *Instituto Superior Tecnico, Portugal*
Sue Dumais, *Microsoft Research*
William Eddy, *Carnegie Mellon University*
Charles Elkan, *University of California, San Diego*
Christos Faloutsos, *Carnegie Mellon University*
Tom Fawcett, *Bell Atlantic Science and Technology*
Usama M. Fayyad, *Microsoft Research*
Ronen Feldman, *Bar-Ilan University, Israel*
Stephen Gallant, *Knowledge Stream Partners*
Clark Glymour, *University of California, San Diego*
Moises Goldszmidt, *SRI International*
Georges Grinstein, *University of Massachusetts, Lowell*
Dimitrios Gunopulos, *IBM Almaden Research Center*
Jiawei Han, *Simon Fraser University, Canada*
David Hand, *Open University, UK*
David Heckerman, *Microsoft Research*
Tomas Imielinski, *Rutgers University*
Yannis Ioannidis, *University of Athens, Greece and University of Wisconsin, Madison*
H.V. Jagadish, *AT&T Laboratories*
David Jensen, *University of Massachusetts, Amherst*
George H. John, *Epiphany*
Pete Johnson, *Mellon Bank Strategic Technology*
Michael Jordan, *Massachusetts Institute of Technology*
Daniel Keim, *University of Halle-Wittenberg, Germany*
Willi Kloesgen, *GMD, Germany*
Ronny Kohavi, *Silicon Graphics*
Hans-Peter Kriegel, *University of Munich, Germany*
T.Y. Lin, *San Jose State University*
Peter Lockemann, *Universitaet Karlsruhe, Germany*
Hongjun Lu, *National University of Singapore, Singapore*
Neil Mackin, *WhiteCross Data Exploration*
David Madigan, *University of Washington*
Heikki Mannila, *University of Helsinki, Finland*
Brij Masand, *GTE Laboratories*
Gary McDonald, *General Motors Global Research and Development Operations*
Eric Mjolsness, *Jet Propulsion Laboratory*
Sally Morton, *Rand Corporation*
Richard Muntz, *University of California, Los Angeles*
Raymond Ng, *University of British Columbia, Canada*
Shojiro Nishio, *Osaka University, Japan*
Ismail Parsa, *Epsilon*
Gregory Piatetsky-Shapiro, *Knowledge Stream Partners*
Daryl Pregibon, *AT&T Laboratories*
Foster Provost, *Bell Atlantic Science and Technology*
Raghu Ramakrishnan, *University of Wisconsin, Madison*
Patricia Riddle, *University of Auckland, New Zealand*
Ted Senator, *National Association of Securities Dealers*
Jude Shavlik, *University of Wisconsin, Madison*
Wei-Min Shen, *University of Southern California*
Kyusoek Shim, *Bell Laboratories*
Arno Siebes, *CWI, The Netherlands*
Evangelos Simoudis, *IBM*
Padhraic Smyth, *University of California, Irvine*
Ramakrishnan Srikant, *IBM Almaden Research Center*
Salvatore J. Stolfo, *Columbia University*
Paul Stolorz, *Jet Propulsion Laboratory*
Kurt Thearling, *Exchange Applications*
Hannu Toivonen, *University of Helsinki, Finland*
Shalom Tsur, *Hitachi America*
Michael Turmon, *Jet Propulsion Laboratory*
Alexander Tuzhilin, *New York University*
Jeffrey D. Ullman, *Stanford University*
Ramasamy Uthurusamy, *General Motors Corporation*
Graham Williams, *CSIRO Mathematical and Information Sciences, Australia*
David Wolpert, *NASA Ames Research Center*
Jan M. Zytkow, *University of North Carolina*

Thursday

Plenary Session

Westside Ballroom South, Fifth Floor

1:30 - 2:00 PM
Opening Remarks
Gregory Piatetsky-Shapiro (General Chair)

Conference Overview
Rakesh Agrawal and Paul Stolorz (Program Chairs)

Best Paper Awards
Gregory Piatetsky-Shapiro and Foster Provost (Publicity Chair)

Paper Track

Westside Ballroom South, Fifth Floor

2:00 - 3:00 PM
Invited Talk
Business Intelligence
Don Haderle

3:00 - 3:30 PM
Report from the Interface Conference
John Elder

3:30 - 4:00 PM
Report from the VLDB Conference

4:00 - 4:30 PM
Coffee Break

4:30 - 6:00 PM
Paper Session 1

Classification

Occam's Two Razors: The Sharp and the Blunt
Pedro Domingos, Instituto Superior Técnico

Interpretable Boosted Naïve Bayes Classification
Greg Ridgeway, David Madigan, Thomas Richardson and John O'Kane, University of Washington

CLOUDS: A Decision Tree Classifier for Large Datasets
Khaled Alsabti, Syracuse University; Sanjay Ranka, University of Florida; Vineet Singh, Hitachi America, Ltd.

Tutorial Program

2:00 - 4:00 PM
Tutorial 1
Westside Ballroom North, Fifth Floor
Database Methods for Data Mining
Heikki Mannila, University of Helsinki, Finland

4:00 - 4:30 PM
Coffee Break

4:30 - 6:30 PM
Tutorial 2
Westside Ballroom North, Fifth Floor
Data Reduction
H. V. Jagadish, AT&T Laboratories and Christos Faloutsos, Carnegie Mellon University and University of Maryland, College Park

6:30 - 7:30 PM
Welcoming Reception
Broadway Ballroom, Sixth Floor

Friday

Tutorial Program

8:00 - 10:00 AM

Tutorial 3
Westside Ballroom South, Fifth Floor

A Tutorial Introduction to High Performance Data Mining

Robert Grossman, Magnify, Inc. and National Center for Data Mining, University of Illinois at Chicago and Stuart Bailey, National Center for Data Mining, University of Illinois at Chicago

Tutorial 4
Westside Ballroom North, Fifth Floor

Fraud Detection and Discovery

Steven K. Donoho and Scott W. Bennett, SRA International, Inc.

Tutorial 5
Marquis Ballroom, Ninth Floor

New-Wave Nonparametric Regression Methods for KDD

David Banks and Mark S. Levenson, National Institute of Standards and Technology

10:00 - 10:30 AM

Coffee Break

10:30 AM - 12:30 PM

Tutorial 6
Marquis Ballroom, Ninth Floor

Smoothing Methods for Learning from Data

J. S. Marron, University of North Carolina

Tutorial 7
Westside Ballroom North, Fifth Floor

Evaluating Knowledge Discovery and Data Mining

Foster Provost, Bell Atlantic Science and Technology and David Jensen, University of Massachusetts, Amherst

Tutorial 8
Westside Ballroom South, Fifth Floor

A Comparison of Leading Data Mining Tools

John F. Elder IV and Dean W. Abbott, Elder Research

12:30 - 2:00 PM

Lunch Break

Exhibit/Demo Program

2:00 - 6:00 PM
Astor Ballroom, Seventh Floor

Exhibits/Demos Open

4:30 - 5:15 PM

Exhibit Talk
Duffy/Columbia, Seventh Floor

Data Mining in the Real World

Gordon Linoff, Data Miners

This presentation will discuss the issues of data mining in the real world. It will touch on the relationship of data mining with a data warehouse (is it really easier?) and on issues related to managing data and choosing particular techniques.

Paper Track

Westside Ballroom, Fifth Floor

2:00 - 2:30 PM

Poster Previews 1 (1-12)

See list on page 7.

2:30 - 3:30 PM

Invited Talk

Mining the World Wide Web

Tom Mitchell, Carnegie Mellon University

3:30 - 4:00 PM

Poster Previews 2 (13-24)

See list on page 7.

4:00 - 4:30 PM

Coffee Break

4:30 - 6:00 PM

Paper Session 2

Associations Rules

Interestingness-Based Interval Merger for Numeric Association Rules

Ke Wang, Soon Hock William Tay and Bing Liu, National University of Singapore

Similarity of Attributes by External Probes

Gautam Das, University of Memphis; Heikki Mannila and Pirjo Ronkainen, University of Helsinki

Integrating Classification and Association Rule Mining

Bing Liu, Wynne Hsu and Yiming Ma, National University of Singapore

6:00 - 7:30 PM

Dinner Break

7:30 - 9:30 PM

Poster Session 1 (1-24)
Westside Ballroom, Fifth Floor

See list on page 7.

Saturday

Exhibit/Demo Program

10:00 AM - 6:00 PM
Astor Ballroom, Seventh Floor
Exhibits/Demos Open

12:00 - 12:45 PM
Exhibit Talk
Duffy/Columbia, Seventh Floor
Data Mining on the Internet: Overview, Algorithmic Challenges and Applications
Shivakumar Vaithaynathan, IBM Research

The advent of the World Wide Web has caused a dramatic increase in the usage of the Internet. The resulting growth in on-line information combined with the almost chaotic nature of the web necessitates the development of powerful yet computationally efficient algorithms. In this talk I will provide examples of applications where data mining could be applied and then focus on the algorithmic challenges along. I will also discuss some new algorithms and provide some results.

4:30 - 5:15 PM
Exhibit Talk
Duffy/Columbia, Seventh Floor
Data Mining Tools
Ismail Parsa, Epsilon

The data mining tools marketplace is diverse. There are tools that offer broad-based data mining capability, tools aimed at solving the problems of a particular industry, tools combined with a service offering, black-box tools and vendors/tools offering custom solutions such as CRM, campaign management, etc.

We will first segment the data mining tools marketplace. We will then learn how to differentiate between the many data mining tools for the best return on investment. We will finally review the summary results of a real-life data mining tool evaluation case study. Come see what all the hype is all about!

Paper Track

Westside Ballroom, Fifth Floor

8:30 - 9:30 AM
Invited Talk
Mining for Dollars: Opportunities and Challenges
Vasant Dhar

9:30 - 10:00 AM
KDD Cup Presentation

10:00 - 10:30 AM
Coffee Break

10:30 AM - 12:00 PM
Paper Session 3

Clustering

An Efficient Approach to Clustering in Large Multimedia Databases with Noise
Alexander Hinneburg and Daniel A. Keim, University of Halle

Scaling Clustering Algorithms to Large Databases
P. S. Bradley, Usama Fayyad and Cory Reina, Microsoft Research

Coincidence Detection: A Fast Method for Discovering Higher-Order Correlations in Multidimensional Data
Evan W. Steeg, Derek A. Robinson and Ed Willis, Molecular Mining Corporation

12:00 - 12:30 PM
Poster Preview 3 (25-36)
Westside Ballroom, Fifth Floor
See list on page 8.

12:30 - 2:00 PM
Lunch Break

2:00 - 3:00 PM
Panel
Database-Data Mining Coupling
Moderator: Rakesh Agrawal, IBM Almaden Research Center

Panelists: Umeshwar Dayal, Hewlett Packard Research; Surajit Chaudhary, Microsoft Research; Tomasz Imielinski, Rutgers University; Heikki Mannila, University of Helsinki, Finland; and Jaiwei Han, Simon Fraser University, Canada

3:00 - 4:00 PM
Paper Session 4

Theory of KDD

Evaluating Usefulness for Dynamic Classification
Gholamreza Nakhaeizadeh, Daimler-Benz Research and Technology; Charles Taylor, University of Leeds; Carsten Lanquillon, Daimler-Benz Research and Technology

A Belief-Driven Method for Discovering Unexpected Patterns
Balaji Padmanabhan, Leonard N. Stern School of Business and Alexander Tuzhilin, Columbia University

4:00 - 4:30 PM
Coffee Break

4:30 - 5:00 PM
Poster Preview 4 (37-49)
Westside Ballroom, Fifth Floor
See list on page 8.

5:00 - 6:00 PM
Panel: Privacy and Data Mining
Moderator: Ellen Spertus, Mills College and Computer Professionals for Social Responsibility

Panelists: Jason Catlett, Junkbusters Corp.; Dan Jaye, Engage Technologies; and Daryl Pregibon, AT&T Labs

Data mining allows unprecedented opportunities for targeted marketing, which can be seen either as a boon for advertisers and consumers or as an enormous invasion of privacy. This panel considers ownership of personal information and explores how maximum benefits can be obtained from data mining while respecting individuals' privacy.

7:30 - 9:30 PM
Cosponsored by SAS Institute Inc.
Discovery Reception and Poster Session 2 (25-49)
Broadway Ballroom North, Sixth Floor
See list on page 8.

Sunday & Monday

Paper Track
Westside Ballroom, Fifth Floor

8:30 - 10:00 AM

Paper Session 5

Discovery in Time

Algorithms for Characterization and Trend Detection in Spatial Databases
Martin Ester, Alexander Frommelt, Hans-Peter Kriegel and Jörg Sander, University of Munich

Pattern Directed Mining of Sequence Data
Valery Guralnik, Duminda Wijesekera and Jaideep Srivastava, University of Minnesota

Rule Discovery from Time Series
Gautam Das and King-Ip Lin, University of Memphis; Heikki Mannila, University of Helsinki; Gopal Renganathan, Autozone Inc.; Padhraic Smyth, University of California, Irvine

10:00 - 10:30 AM
Coffee Break

10:30 - 11:30 AM

Paper Session 6

Applications A

Data Mining for Direct Marketing: Problems and Solutions
Charles X. Ling and Chenghui Li, The University of Western Ontario

A Data Mining Support Environment and Its Application on Insurance Data
M. Staudt, J.-U. Kietz and U. Reimer, Swiss Life

11:30 AM - 12:30 PM

Panel
Behind-the-Scenes Data Mining
Moderator: George H. John, Epiphany

Panelists: Graham Spencer, Excite; Gerald Fahner, Fair, Isaac & Co.; and Paul DuBose Analytika, Inc.

Truly successful technologies become invisible. Data mining has a long way to go - or does it? Most stories of data mining applications focus on an expert's use of raw technology to solve one particular problem, but millions of people use or are affected by data mining technology every day, without even being aware of it. The panelists will discuss examples and characteristics of this "behind-the-scenes" variety of data mining.

12:30 - 2:00 PM
Lunch Break

Luncheon meeting
KDD-98 Program Committee Meeting
Odets/Wilder, Fourth Floor

2:00 - 3:00 PM

Paper Session 7

Applications B

Finding Frequent Substructures in Chemical Compounds
Luc Dehaspe, Katholieke Universiteit Leuven; Hannu Toivonen, University of Helsinki; Ross Donald King, The University of Wales, Aberystwyth

Mining Audit Data to Build Intrusion Detection Models
Wenke Lee, Salvatore J. Stolfo and Kui W. Mok, Columbia University

3:00 - 4:00 PM
Open Feedback/Planning Session

Monday, August 31

Workshop Program

By invitation only

9:00 AM - 5:00 PM

KW1
Lyceum Complex, Fifth Floor
Data Mining in Finance
Chairs: Tae Horn Hann, University of Karlsruhe, Germany and Gholamreza Nakhaeizadeh, Daimler-Benz AG, Research and Technology, Germany

9:00 AM - 5:00 PM

KW2
Odets/Wilder, Fourth Floor
Distributed Data Mining
Chairs: Hillol Kargupta, Washington State University and Philip Chan, Florida Institute of Technology

9:00 AM - 5:00 PM

KW3
Julliard Complex, Fifth Floor
Keys to the Commercial Success of Data Mining
Chairs: Kurt Thearling, Exchange Applications and Roger M. Stein, Moody's Investors Service

Friday Poster Sessions

Poster Session 1

7:30 - 9:30 PM
Westside Ballroom, Fifth Floor
(Poster Previews: 2:00 - 2:30 PM and 3:30 - 4:00 PM, Westside Ballroom, Fifth Floor)

Posters 1-12

Online Generation of Profile Association Rules
Charu C. Aggarwal, T. J. Watson Research Center; Zheng Sun, Duke University; Philip S. Yu, T. J. Watson Research Center

ADtrees for Fast Counting and for Fast Learning of Association Rules
Brigham Anderson and Andrew Moore, Carnegie Mellon University

Independence Diagrams: A Technique for Visual Data Mining
Stefan Berchtold and H. V. Jagadish, AT&T Laboratories; Kenneth A. Ross, Columbia University

Direct Marketing Response Models Using Genetic Algorithms
Siddhartha Bhattacharyya, University of Illinois at Chicago

Mining Association Rules in Hypertext Databases
José Borges and Mark Levene, University College London

Blurring the Distinction between Command and Data in Scientific KDD
John Carlis, Elizabeth Shoop and Scott Krieger, University of Minnesota

Probabilistic Modeling for Information Retrieval with Unsupervised Training Data
Ernest P. Chan, Credit Suisse First Boston; Santiago Garcia, Morgan Stanley & Co. Inc.; Salim Roukos, IBM T. J. Watson Research Center

Toward Scalable Learning with Non-Uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection
Philip K. Chan, Florida Institute of Technology and Salvatore J. Stolfo, Columbia University

Joins that Generalize: Text Classification Using WHIRL
William W. Cohen, AT&T Labs—Research and Haym Hirsh, Rutgers University

Giga-Mining
Corinna Cortes and Daryl Pregibon, AT&T Labs—Research

Interactive Interpretation of Kohonen Maps Applied to Curves
Anne Debregeas and Georges Hebrail, Electricite de France

FlexiMine—A Flexible Platform for KDD Research and Application Construction
C. Domshlak, D. Gershkovich, E. Gudes, N. Liusternik, A. Meisels, T. Rosen and S. E. Shimony, Ben-Gurion University

Posters 13-24

A Fast Computer Intrusion Detection Algorithm Based on Hypothesis Testing of Command Transition Probabilities
William DuMouchel, AT&T Labs—Research and Matthias Schonlau, AT&T Labs—Research and National Institute of Statistical Sciences

Initialization of Iterative Refinement Clustering Algorithms
Usama Fayyad, Cory Reina and P. S. Bradley, Microsoft Research

Mining in the Presence of Selectivity Bias and Its Application to Reject Inference
A. J. Feelders, Tilburg University; Soong Chang and G. J. McLachlan, University of Queensland

On the Efficient Gathering of Sufficient Statistics for Classification from Large SQL Databases
Goetz Graefe, Usama Fayyad and Surajit Chaudhuri, Microsoft Corporation

Coactive Learning for Distributed Data Mining
Dan L. Grecu and Lee A. Becker, Worcester Polytechnic Institute

Mining Segment-Wise Periodic Patterns in Time-Related Databases
Jiawei Han, Wan Gong and Yiwen Yin, Simon Fraser University

Learning to Predict the Duration of an Automoblie Trip
Simon Handley and Pat Langley, Daimler-Benz Research and Technology Center; Folke A. Rauscher, Daimler-Benz AG

Fast Computation of Two-Dimensional Depth Contours
Ted Johnson, AT&T Research Center; Ivy Kwok and Raymond Ng, University of British Columbia

Comparing Massive High-Dimensional Data Sets
Theodore Johnson and Tamraparni Dasu, AT&T Labs—Research

Defining the Goals to Optimise Data Mining Performance
Mark G. Kelly, David J. Hand and Niall M. Adams, The Open University

An Enhanced Representation of Time Series which Allows Fast and Accurate Classification, Clustering and Relevance Feedback
Eamonn J. Keogh and Michael J. Pazzani, University of California, Irvine

Active Templates: Comprehensive Support for the Knowledge Discovery Process
Randy Kerber, Hal Beck, Tej Anand and Bill Smart, NCR Human Interface Technology Center

Saturday Poster Sessions

7:30 - 9:30 PM

Poster Session 2

Broadway Ballroom North, Sixth Floor

(Poster Previews: 12:00 - 12:30 PM and 4:30 - 5:00 PM, Westside Ballroom, Fifth Floor)

Posters 25-36

Targeting Business Users with Decision Table Classifiers
Ron Kohavi and Daniel Sommerfield, Silicon Graphics, Inc.

BAYDA: Software for Bayesian Classification and Feature Selection
Petri Kontkanen, Petri Myllymäki, Tomi Silander and Henry Tirri, University of Helsinki

Approaches to Online Learning and Concept Drift for User Identification in Computer Security
Terran Lane and Carla E. Brodley, Purdue University

Human Performance on Clustering Web Pages: A Preliminary Study
Sofus A. Macskassy, Arunava Banerjee, Brian D. Davison and Haym Hirsh, The State University of New Jersey

Aggregation of Imprecise and Uncertain Information for Knowledge Discovery in Databases
Sally McClean, Bryan Scotney and Mary Shapcott, University of Ulster

Discovering Predictive Association Rules
Nimrod Megiddo and Ramakrishnan Srikant, IBM Almaden Research Center

Reinforcement Learning for Trading Systems and Portfolios
John Moody and Matthew Saffell, Oregon Graduate Institute

Group Bitmap Index: A Structure for Association Rules Retrieval
Tadeusz Morzy and Maciej Zakrzewicz, Poznan University of Technology

Towards Personalization of Algorithms Evaluation in Data Mining
Gholamreza Nakhaeizadeh, Daimler-Benz AG and Alexander Schnabl, Technical University Vienna

Large Datasets Lead to Overly Complex Models: An Explanation and a Solution
Tim Oates and David Jensen, University of Massachusetts

Analysing Rock Samples for the Mars Lander
Jonathan Oliver, University of California, Berkeley; Ted Roush and Paul Gazis, NASA Ames Research Center; Wray Buntine, Rohan Baxter and Steve Waterhouse, Ultimode Systems

Memory Placement Techniques for Parallel Association Mining
Srinivasan Parthasarathy, Mohammed J. Zaki and Wei Li, University of Rochester

Posters 37-49

Methods for Linking and Mining Massive Heterogeneous Databases
José C. Pinheiro and Don X. Sun, Bell Laboratories

Mining Databases with Different Schemas: Integrating Incompatible Classifiers
Andreas L. Prodromidis and Salvatore Stolfo, Columbia University

Time Series Forecasting from High-Dimensional Data with Multiple Adaptive Layers
R. Bharat Rao, Scott Rickard and Frans Coetzee, Siemens Corporate Research, Inc.

Ranking-Methods for Flexible Evaluation and Efficient Comparison of Classification Performance
Saharon Rosset, Tel Aviv University

A Robust System Architecture for Mining Semi-Structured Data
Lisa Singh, Bin Chen, Rebecca Haight, Peter Scheuermann and Kiyoko Aoki, Northwestern University

Defining *diff* as a Data Mining Primitive
Ramesh Subramonian, Intel Corporation

Simultaneous Reliability Evaluation of Generality and Accuracy for Rule Discovery in Databases
Einoshin Suzuki, Yokohama National University

Mining Generalized Association Rules and Sequential Patterns Using SQL Queries
Shiby Thomas, University of Florida and Sunita Sarawagi, IBM Almaden Research Center

Data Reduction Based on Hyper Relations
Hui Wang, Ivo Düntsch and David Bell, University of Ulster

Discovering Technical Traders in the T-bond Futures Market
Andreas S. Weigend, Fei Chen and Stephen Figlewski, New York University; Steven R. Waterhouse, Ultimode Systems

Learning to Predict Rare Events in Event Sequences
Gary M. Weiss, AT&T Labs and Rutgers University and Haym Hirsh, Rutgers University

Daily Prediction of Major Stock Indices from Textual WWW Data
B. Wüthrich, D. Permunetilleke, S. Leung, V. Cho, and J. Zhang, The Hong Kong University of Science and Technology; W. Lam, The Chinese University of Hong Kong

PLANMINE: Sequence Mining for Plan Failures
Mohammed J. Zaki, Neal Lesh and Mitsunori Ogihara, University of Rochester

Exhibitors

AAAI Press

445 Burgess Drive
Menlo Park, CA 94035
Tel.: 650-328-3123
Fax: 650-321-4457
E-mail: press@aaai.org
Web: www.aaai.org/Press/

AAAI Press has published books about KDD since 1989. We will be exhibiting the popular *Advances in Knowledge Discovery* (Fayyad et al.) as well as back issues of the KDD conference proceedings, technical reports in KDD, and other fine books in AI and computer science. Conference attendees may take a 20% discount on all on-site book orders!

AZMY Thinkware, Inc.

Contact: Hany Azmy
1450 Palisade Ave. #M1D
Fort Lee, NJ 07024
Tel.: 201-947-1881
Fax: 201-947-1804
E-mail: mail@azmy.com
Web: www.azmy.com

AZMY Thinkware is exhibiting SuperQuery 2.0 Discovery Edition, a data analysis and mining tool that runs under Windows 95 and NT. Using rule induction technology, SuperQuery searches data tables and reports all interesting patterns and exceptions. The Fact Discovery Engine is easily tuned to meet various analysis needs. SuperQuery is the only tool that allows "remining" of the discovered facts.

SuperQuery also assists in preparing data for analysis by providing a number of facilities for partitioning, classifying and processing data columns. In addition, SuperQuery helps users explore and analyze their data by automatically displaying graphs and calculating statistics. It contains a number of Wizards that help read, update, and analyze data effortlessly. SuperQuery can access and query a number of databases, spreadsheets, text files directly, and through ODBC drivers. A 16-bit version of SuperQuery is also available for Windows 3.x. SuperQuery is used in many applications including quality control, survey analysis, medical studies, and defense. Bring a sample of your data to our booth and we will show you what SuperQuery can discover for you.

Compression Sciences Limited

Contact: Miranda Noy
2 Chalfont Square, Old Foundry Road
Ipswich, Suffolk IP4 2AJ UK
Tel.: 44-1473-267103
Fax: 44-1473-267104
E-mail: mnoy@gentia.com

For knowledge discovery (KD) software to succeed, the enthusiasm of the innovator and early adopter customers must be translated into practical benefits for the majority. This demands that KD software be made easy to use. Users familiar with Web browsers must be able to navigate and manipulate KD projects with ease. The systems must run from the Web, run on mainstream client server hardware and address very large data volumes.

K.wiz from Compression Sciences delivers a new KD solution. Combining ease of use with Scalability, this advanced client server framework encompasses all nine stages of the KD process. K.wiz components within this frame deliver data transformation, visualization, and discovery algorithms to the desk and Web top. External Components extend the already powerful range of functionality and ensure each organization's unique demands are leveraged by K.wiz.

Designed for ease of use, K.wiz provides wizards for the novice user and expert mode for the experienced knowledge worker. A full range of Automation, Scheduling Agents and API's empower the application developer to capture K.wiz plans and embed them in custom applications.

Compression Sciences are showcasing K.wiz at KDD-98. Due for launch this Fall, information and demonstrations are available at the Compression Sciences booth.

Data Mining Technologies Inc.

Contact: Michael Gilman
1500 Hampstead Turnpike
East Meadow, NY 11554
Tel.: 516-542-8900
E-mail: info@data-mine.com
Web: www.data-mine.com

Data Mining Technologies Inc. has developed a unique new data mining technology and embedded it in a data mining toolkit called Nuggets™.

Nuggets™ uses state of the art, proprietary computer algorithms to search databases for patterns in the form of rules. It differs from other rule induction methods in that it is not statistically based and therefore does not require any statistical assumptions. It handles missing and noisy data. Other tree building methods build rules by looking at one attribute at a time. This means that complex non-linear interactions among variables that are the essence of the power of data mining might be overlooked. Nuggets™ however is a true rule induction system which searches for simultaneous attribute interactions. This insures that *all* rules are implicitly searched, thereby releasing the full power of data mining.

Nuggets™ predicts, classfies, segments and validates and is easy to use. Ask about the many unique features to simplify your data mining needs.

Elsevier Science

Contact: Traci Taylor
655 Avenue of the Americas
New York, NY 10010
Tel.: 212-633-3766
Fax: 212-633-3764
E-mail: t.taylor@elsevier.com
Web: www.elsevier.com

ISL Decision Systems Inc.

630 Freedom Business Center, Suite 314
King of Prussia, PA 19406
Tel.: 610 768 7725
Fax: 610 768 7774
Web: http://www. isldsi.com

ISL Decision Systems Inc. is a leading data mining software company. It is part of the ISL Group which also has affiliates in the U.K. and Singapore and a network of distributors serving countries around the world. Launched in 1994, the Group's award winning Clementine product was the first enterprise -strength data mining system to be aimed at business users. There are now over 550 clients worldwide.

Clementine is consistently acknowledged by users and analysts as the leading interactive data mining system and the ISL Group has led the way in discovering new applications for data mining including fraud prevention, pharmaceutical research, consumer buying patterns, financial risk assessment, point of sale data analysis and customer profiling and production process analysis.

Current developments include a collaboration with NCR and Daimler Benz on an open standard methodology for data mining, new middleware to provide truly scalable data mining performance on multiple database platforms and mining of the web, on the web.

ISOFT

Contact: Raphaelle Thomas
Chemin du Moulon
91190 Gif sur Yvette
France
E-mail: rthomas@isoft.fr
Web: www.isoft.fr

ALICE D'ISOFT is a powerful desktop data mining tool designed for mainstream business users. Based on decision tree technology, its user-friendly interface and visual data mining approach makes it ideal for marketing managers, commercial directors, financial directors and all decision makers who need to make strategic decisions.

ALICE D'ISOFT accesses company data bases directly, segments and classifies the data and allows decision makers to test their hypotheses. That means answers to questions such as: Who is the target audience for my product? Which type of clients represents a credit risk? etc.

Applications areas: population analysis, risk evaluation, data classification, forecasting, quality control.

Kluwer Academic Publishers

Contact: Marcia Kidston
101 Philip Drive
Norwell, MA 02061
Tel.: 781-871-6600
Fax: 781-871-6528
E-mail: kluwer@wkap.com
Web: www.wkap.nl

Exhibitors

McGraw-Hill Publishers

Contact: Mary Jo Donnelly
WCB/McGraw-Hill
1333 Burr Ridge Parkway
Burr Ridge, IL 60521
Tel.: 800-634-3963

McGraw-Hill is a leading publisher of computer science texts. We continue our excellence with titles such as *Machine Learning* by Tom Mitchell, *Database Management Systems* by Raghu Ramakrishnan, and *Database System Concepts* by Abraham Silberschatz, Henry Korth and S. Sudarshan. We invite you to stop by and browse through our catalog and book display, as we will be offering discounts between 10%-30% at this conference.

Megaputer Intelligence

Contact: Sergei Ananyan
1518 E. Fairwood Drive
Bloomington, IN 47408
Tel.: 812-339-1646
Fax: 812-339-1646
E-mail: megaputers@aol.com
Web: www.megaputer.ru

PolyAnalyst is a unique data mining solution for Windows NT and 95. PolyAnalyst is a complete multi-strategy data mining environment based on the latest achievements in the field of automated knowledge discovery in databases. PolyAnalyst presents the discovered relations in explicit symbolic form. A large selection of exploration engines allows the user to predict values of continuous variables, model complex phenomena, determine the most influential independent variables, and solve classification and clustering tasks. An object-oriented design, point-and-click GUI, versatile visualization and reporting capabilities, minimum of statistics, and a simple interface with data storage architectures make PolyAnalyst a very easy-to-use system.

PolyAnalyst for Windows NT has a solid record of successful applications in marketing, banking, finance, insurance, retailing, and pharmaceuticals. The system is also available in the Client/Server architecture, while a simplified system, PolyAnalyst Lite, works under Windows 95. A free evaluation copy of the software is available at http://www.megaputer.ru

"Unlike neural network programs, PolyAnalyst displays a symbolic representation of the relationship between the independent and dependent variables. This is a critical advantage for business applications, because managers are reluctant to use a model if they don't understand how it works," says Raymond Burke, Kelley Chair of BA at IU.

Morgan Kaufmann Publishers

Contact: Katja Kolinke
340 Pine Street, 6th Floor
San Francisco, CA 94104-3205
Tel.: 415-392-2665
Fax: 415-982-2665
E-mail: mkp@mkp.com
Web: www.mkp.com

Morgan Kaufmann publishes the finest technical information resources for computer science and engineering professionals. We publish in book and digital form in such areas as databases, networking, computer architecture, human computer interaction, computer graphics, multimedia information systems, artificial intelligence, and software engineering. Many of our books are considered to be the definitive works in their fields.

We believe strongly in seeking out the most authoritative, expert authors. Our family of authors and series editors includes many of the world's most respected computer scientists and engineers and their books often represent the wisdom gained from years of research, development, and teaching.

We believe it is our responsibility to add value to our books by working with authors to improve content and exposition. MK books are extensively peer reviewed and, in the case of textbooks, are often class tested with hundreds of students. All of our books are professionally edited.

PC AI Magazine

Contact: Robin Okun
P.O. Box 30130
Phoenix, AZ 85046
Tel.: 602-971-1869
Fax: 602-971-2321
E-mail: info@pcai.com
Web: www.pcai.com/pcai/

PC AI Magazine provides the information necessary to help managers, programmers, executives, and other professionals understand the quickly unfolding realm of artificial intelligence (AI) and intelligent applications (IA). PC AI addresses the entire range of personal computers including the Mac, IBM PC, neXT, Apollo, and more. PC AI features developments in expert systems, neural networks, object oriented development, and all other areas of artificial intelligence. Feature articles, product reviews, real-world application stories, and a Buyer's Guide present a wide range of topics in each issue.

Salford Systems

Contact: Kerry Martin
Salford Systems
8880 Rio San Diego Drive, Suite 1045
San Diego, CA 92108
Tel.: 619-543-8880
Fax: 619-543-8888
E-mail: info@salford-systems.com
Web: www.salford-systems.com

CART is a robust and scalable decision-tree tool for data mining, predictive modeling and data preprocessing. CART automatically discovers cause-and-effect relationships, isolates significant patterns and forecasts trends. The software's advanced functionality and new resampling technology, deployed via a highly intuitive graphical user interface, generates accurate and reliable predictive trees that graphically depict which factors drive the results and how.

As an affordable, stand-alone application, CART's unique combination of automated solutions-including adjustable misclassification penalties, embedded self-validation procedures, committees of experts, and missing-value surrogates-empowers business users and data analysts to effectively tackle real-world modeling problems. And, when used as powerful supplemental analysis, CART improves the performance of other data-mining techniques (e.g., neural networks and logistic regression).

Worldwide, CART has more than 1,000 users found in nearly all industry segments, including marketing, financial services, insurance, retail, healthcare, pharmaceutical, manufacturing, telecommunications, energy, agricultural, and education. In these data-intensive industries, CART is especially efficient harvesting a high return on companies' investments in large, complex data warehouses. CART applications span market research segmentation, direct marketing, fraud detection, credit scoring, risk management, biomedical research and manufacturing quality control. Users include AT&T Universal Card Services, Cabela's, Fleet Financial Group, Pfizer Inc., and Sears, Roebuck and Co.

Exhibitors

SAS Institute Inc.

Contact: Rich Rovner
SAS Campus Drive
Cary, NC 27513
Tel.: 919-677-8000
Fax: 919-677-4444
Web: www.sas.com

SAS Institute Inc., the world's largest privately held software company and a leader in data mining software, is presenting Enterprise Miner, a complete process-driven solution for small- and large-scale data mining applications. Enterprise Miner builds and extends 22 years of proven analytic software into a complete, GUI-based solution integrating traditional statistics, computational methods, and AI. Enterprise Miner provides regression, neural networks, decision trees, clustering, associations, sequences, visualization, transformation, outlier handling, assessment methods, automatic scoring, model manager, and a modeling API. An interactive process flow interface presents this and more, allowing statisticians and analysts to develop, assess, and share solutions.

Enterprise Miner can be deployed in a client/server environment for fully scalable processing, and output is Web enabled, delivering results organization-wide. Yphise, the respected firm of industry anaysts specializing in software evaluations, awarded Enterprise Miner top marks in its survey of data mining solutions.

In addition to software, SAS Institute offers training and consulting services to provide numerous paths to data mining expertise. Courses are available on data mining techniques and Enterprise Miner usage. SAS consultants can work on short- or long-term data mining projects, delivering solutions and knowledge transfer.

Sentient Machine Research

Contact: Peter van der Putten
Baarsjesweg 224 1058 AA
Amsterdam
The Netherlands
Tel.: 31-20-6186927
Fax: 31-20-6124504
E-mail: info@smr.nl
Web: www.smr.nl

Sentient Machine Research, a Dutch R&D company founded in 1990, develops software and technology in image processing, multimedia information retrieval and data mining. At KDD-98 will present the new 2.0 release of its successful DataDetective datamining environment. DataDetective is build around an efficient fuzzy search and match engine and covers the full range of data mining functionalities: predictive modeling, profiling, clustering and segmentation. (The model generated by DataDetective's modeling assistant 'Targas' ended at a respectable fifth place in last years KDD-Cup) Most distinctive within the DataDetective environment is the graphical clustering tool 'Looking Glass,' based on proprietary animated clustering algorithms, which allows for a uniquely interactive style of visual data mining.

Silicon Graphics

Contact: Aydin Senkut
2051 N. Shoreline Blvd.
Mail Stop 08L-855
Mountain View, CA 94043

MineSet™ is Silicon Graphics' industry leading integrated data mining product. Mineset offers a unique combination of scalable performance, intuitive user interface, unparalleled visualization features and sophisticated analytics, geared towards both technical and business users.

The Meta Group ranked MineSet third (behind statistical packages) in data mining market share in its January 1998 industry report on data warehouse marketing trends and opportunities. MineSet also won the Bronze Miner Award and ranked highest among commercial data mining vendor products in last August's knowledge discovery and data mining (KDD) competition.

MineSet 2.5 provides the user with a revolutionary paradigm for knowledge discovery by offering parallelized data mining algorithms for faster performance as well as new analytical tools, such as regression, clustering, and decision tables for more intuitive comprehension of data. Combining powerful integrated, interactive tools for data access and transformation, analytical data mining, and visual data mining, MineSet will maximize the value of your data.

SRA International

Contact: Jim Hayden
4350 Fair Lakes Court
Fairfax, VA 22033
Tel.: 703-503-1856
Fax: 703-803-1509
E-mail: jim_hayden@sra.com
Web: www.sra.com

SRA International, Inc. offers a complete line of fully scaleable data mining tools and professional services, empowering organizations with the ability to discover and detect patterns critical to their success.

SRA's KDD Explorer toolset includes multistrategy algorithms for discovering associations, classifications, sequences, and clusters, as well as high-speed rule and sequence-based pattern matching algorithms. These algorithms access relational databases directly for mining data, using parallel computing methods to exploit powerful multiprocessor platforms and rapidly analyze extremely large data sets. Our user interfaces are JDBC-compliant and Java-based, communicating to the RDBMS across distributed networks. KDD Explorer offers serious data mining professionals an integrated workflow-driven environment for configuration and execution of algorithms as well as visualization or results for analysis and interpretation.

SRA's knowledge discovery specialists understand how best to apply these advanced capabilities to enable you to utilize your most strategic asset: electronic information. Together, SRA's KDD Explorer toolset and professional services provide solutions giving you flexibility and power to apply to business areas such as fraud detection and prevention, cost understanding, competitive intelligence, and trend analysis.

SRA International has been creating innovative solutions to practical problems faced by businesses and government agencies for twenty years. We specialize in the fields of intelligent information retrieval; machine learning; knowledge-based systems; database engineering; and natural language processing.

Tektonic Software, Tandem Division of Compaq

Contact: Dee Dobbs
10400 N. Tantau Ave. #248-49
Cupertino, CA 95014
Tel.: 408-285-3280
Fax: 408-285-3255
E-mail: dee.dobbs@tandem.com
Web: www.tektonic.com

The InfoCharger's speed enables OLAP or data mining tools to work on large volumes of data. InfoCharger is a software component for interactive sessions on inexpensive hardware. The InfoCharger allows users to process detailed data instead of working on condensed subtotals, enabling them to discover meaningful patterns. This kind of data analysis is critical in order to make better decisions or exploit new business opportunities, including increased sales productivity and cost controls through more effective target marketing campaigns; new growth opportunities by identification of cross-selling possibilities; reduced exposure to risk or fraud; effective churn management; and optimized resource allocation

To exploit the full spectrum of this market, Tektonic Software is partnering with tool vendors, specialists in vertical markets and application consultants.

Thinking Machines Corporation

Contact: Charles Berger
16 New England Executive Park
Burlington, MA 01803
Tel.: 781-238-3418
Fax: 781-238-3440
E-mail: cberger@think.com
Web: www.think.com

Darwin is an easy-to-use, Windows client/UNIX server, scalable, multi-algorithmic data mining software suite designed to build predictive models from large customer databases. Darwin supports prediction and classification modeling via neural networks, classification and regression tress (C&RT), and k-nearest neighbor algorithms.

As an open solution, Darwin can run on some of the world's fastest and most powerful comput-

Exhibitors

ing platforms running Sun Solaris and HP-UX including parallel processing, and symmetric multiprocessors (SMP) configurations. With Darwin, financial services, telecommunications, database marketing companies and other large corporations can uncover vital information that was previously undetected because of the sheer size and complexity of their database.

Darwin also offers a scripting tool that records data mining steps to re-run and automate the data mining process and a workflow feature that graphically documents the data mining steps and provides information about each of the steps taken.

An optional feature of Darwin is the ability to generate predictive models in C, C++ or Java code for deployment outside of the Darwin environment. These deployable models can be easily integrated into existing systems and procedures, so any new business information can be made available where and when it's most needed-for example, in call centers and web-based applications.

Ultimode Systems

Contact: Steve Waterhouse
2560 Bancroft Way #213
Berkeley, CA 94704
Tel.: 510-548-8978
Fax: 510-845-2292
E-mail: stevew@ultimode.com
Web: www.ultimode.com

ACPro is an data mining tool for automatic segmentation of databases. Segmentation (or clustering) is the discovery of similar groups of records in databases. ACPro is based on the successful NASA AutoClass research program, and was developed in collaboration with the AutoClass team using a NASA commercialization award.

Unlike other segmentation tools, ACPro discovers the optimal number of segments without requiring user specification. It also is an order of magnitude faster than its predecessor, AutoClass. ACPro handles missing data in a coherent manner. It also assigns relevance to the attributes in each segment which aids understanding.

ACPro has been applied in a number of commercial and scientific settings. It has been used for analysis of telecommunications churn data, market segmentation, visualization of geological data and is currently being used at NASA for spectral analysis of rock samples.

ACPro is available with either a command line or platform independent GUI interface for Windows NT/95 and most Unix platforms.

Unica Technologies

Contact: Scott Sassone
Lincoln North
Lincoln, MA 01773-1125
Tel.: 781-259-5900
Fax: 781-259-5901
E-mail: ssassone@unica-usa.com
Web: www.unica-usa.com

Unica Technologies, Inc. is the leading provider of data mining and predictive modeling software and services for database marketing and customer relationship optimization. MODEL 1, our award-winning product line for marketing applications, includes templates for response modeling, cross selling, customer segmentation, and customer valuation. PRW (Pattern Recognition Workbench) our general purpose data mining tool offers twelve algorithms, six methods of intelligent automation and optimization, and three methods of validation.

Both PRW and Model 1 are scalable, from the desktop up to NT and Unix SMP servers. These fully integrated applications offer everything needed for data access, pre-processing, modeling, results interpretation and deployment. API's are also available for customized applications. Unica wrote the book on data mining. Our textbook *Solving Data Mining Problems* is published by Prentice-Hall.

Unica focuses on providing business solutions, not just software. Our Integrated Customer Management Program ensures that Unica's education, training and consulting services are tailored to meet your needs, providing a demonstratable ROI. Put our worldwide experience to work for you today!

Urban Science

Contact: Mark Yuhn
200 Renaissance Center, Suite 1900
Detroit, MI 48243
Tel.: 313-259-9900
Fax: 313-259-1362
E-mail: mcyuhn@urbanscience.com
Web: www.urbanscience.com

*Gain*Smarts Database Marketing Modeling System utilizes sophisticated predictive modeling technology that can analyze past purchase behavior, demographic and lifestyle characteristics, promotion and risk information to predict the likelihood of response as well as to develop an understanding of consumer characteristics.

Economic analysis is then applied to these results, allowing the optimum expenditure of marketing resources to achieve your prospecting, retention or cross-selling objectives. *Gain*Smarts has proven to be extremely effective for businesses worldwide, demonstrating ROI for its users in many application areas.

Additionally, the system received a first place "Gold Miner" award in the KDD-cup 97 competition at the Knowledge Discovery and Data Mining Conference, Newport Beach, CA sponsored by the American Association for Artificial Intelligence (AAAI). The competition involved targeting prospects for a financial services promotion and then comparing predicted vs. actual behavior.

WizSoft

Contact: Abraham Meidan
3 Beit-Hillel St.
Tel-Aviv 67017
ISRAEL
E-mail: info@wizsoft.com
web: www.wizsoft.com

WizSoft exhibits two data mining applications, WizWhy and WizRule. WizWhy is a knowledge discovery application based on a proprietary association rules algorithm. WizWhy reveals all the if-then rules that relate to the dependent variable, and uses these rules in order to issue predictions, summarize the data and reveal unexpected phenomena. WizWhy avoids overfitting by calculating the error probability of each rule.

WizRule is a data auditing application based on data mining technology. WizRule reveals all the if-then rules and the mathematical formulas that govern the data under analysis, and points at the records that deviate from the set of the discovered rules as cases to be audited. WizRule avoids false alarms by calculating the level of unlikelihood of each deviation.

Both products run on Windows 95 / 98 / NT, read any ODBC compliant database, and have OCX versions that can be embedded in other applications.

Demos

BAYDA (Bayesian Discriminant Analysis)

Paper Title: BAYDA: Software for Bayesian Classification and Feature Selection

Development Team: Henry Tirri, Petri Kontkanen, Jussi Lahtinen, Petri Myllymäki, Tomi Silander, University of Helsinki

Tel.: +358-9-708-44173

BAYDA is a Java software package for flexible data analysis in predictive data mining tasks. BAYDA performs fully Bayesian predictive inference of class memberships based on a Naive Bayes model build from the data set. It is well-known that the Naive Bayes classifier performs well in predictive data mining tasks, when compared to approaches using more complex models. However, the model makes strong independence assumptions that are frequently violated in practice. For this reason, the BAYDA software also provides a feature selection scheme which can be used for analyzing the problem domain, and for improving the prediction accuracy of the models constructed by BAYDA. The feature selection can be done either manually or automatically. In manual selection the user has an opportunity to use BAYDA for evaluating different feature subsets by leave-one-out cross validation scheme. In the automatic feature selection case the program selects the relevant features by using a novel Bayesian criterion.

The current version features of BAYDA include (1) missing data handling; (2) an external leave-one-out cross validated estimate of the classifier performance in graphical format; (3) "intelligent document" style graphical interface; (4) forward selection/backward elimination feature subset selection; (5) free format data files (such as tab-delimited format of SPSS).

BAYDA is available free of charge for research and teaching purposes from www.cs.Helsinki.FI/research/cosco under section "Software," and it is currently tested on Windows'95/NT, SunOS and Linux platforms. However, being implemented in 100% Java, it should be executable on all platforms supporting Java Runtime Environment 1.1.3 or later.

What Is Unique about the System? (1) intelligent, adaptive HTML-document interface; (2) Bayesian criterion for variable subset selection; (3) fully Bayesian prediction based on model parameter averaging.

Bayesian Knowledge Discoverer (BKD)

Development Team: Marco Ramoni and Paola Sebastiani, The Open University

Tel.: 413-577-0338

Bayesian Knowledge Discoverer (BKD) computer program to discover of Bayesian belief networks (BBNs) from (possibly incomplete) databases. A BBN is a direct cyclic graph where nodes represent stochastic variables and direct arcs identify dependencies between a set of parent variables and a child variable. Each dependency is then quantified by a conditional probability distribution shaping the behavioral relationships between the set of parent variables and the child variable. In this way, a BBN provides a dependency model of the underlying domain knowledge and a graphical representation of decision problems, grounded on solid foundations of probability theory, able to perform prediction, explanation and classifications. Given a database, BKD is able to extract the graphical structure from data, estimate the conditional probability distributions from data, discretize continuos variables, handle missing data, automatically define network nodes from data. Once generated, the extracted BBN can be used as a self-contained intelligent decision support system able to provide predictions and explanations. A goal-oriented propagation algorithm is included in BKD. A graphical user interface capitalizes on the graphic nature of BBN to allow the user to easily navigate the dependencies embedded in the database. BKD is currently distributed in over 1,000 copies world wide.

What Is Unique about the System? BKD is the first available program implementing a Bayesian approach to the discovery of BBNs. BKD uses a novel method, called Bound and Collapse, to efficiently handle missing data.

Belief Network PowerConstructor 1.0

Development Team: Jie Cheng, University of Ulster

Tel.: +44-1232-366500

Belief network is a powerful knowledge representation and reasoning tool under conditions of uncertainty. In DM systems, it can be used for classification, predication and decision support. Because belief network can handle uncertain information in a natural way and the learned knowledge is well structured and can be easily understood it becomes more and more popular in DM research. To use belief networks in DM systems, a crucial step is to learn the belief networks from large training data sets efficiently and accurately.

PowerConstructor is such a belief network learning tool, which includes a user-friendly interface and a construction engine. The system takes a database table as input and constructs the belief network structure as output. The construction engine is based on our three-phase belief network learning algorithm, which takes an information theoretical approach and has the complexity of $O(N^4)$ on conditional independence (CI) test while all other algorithms require exponential number of CI tests. (N is the number of attributes.) We evaluate our system using a widely accepted benchmark data set with 37 attributes and 10,000 records and other data sets. The results show that our system is the most accurate and efficient system available. The system is available for evaluation at our web site (mmr.infj.ulst.ac.uk/jcheng/bnpc.htm and infosys.susqu.edu/bnpc/) and enjoys over 700 downloads from academic and industrial users. From the encouraging feedback we know that some users have already used it to solve real-world problems.

What Is Unique about the System? 1. Accessibility. It supports most of the desktop database formats and all database servers through ODBC. 2. Reusability. The construction engine is a class library so that it can be easily integrated into DM systems for Windows 95/NT as a component. 3. Efficiency and Accuracy. Both experimental results and theoretical analysis show that our system is better than other systems for belief network learning. 4. User-friendly interface with online help.

DBMiner: A Multi-Functional On-Line Analytical Mining System

Development Team: Jiawei Han, Sonny Chee and Jenny Chiang, Simon Fraser University

Tel.: 604-291-4411 or 604-291-5371

A data mining system, DBMiner, has been developed for interactive mining of multiple-level knowledge in large relational databases and data warehouses. The system implements a wide spectrum of data mining functions, including characterization, comparison, association, classification, prediction, clustering, data dispersion analysis and time-series analysis. It also builds up a user-friendly, interactive data mining environment and a set of knowledge visualization tools. In-depth research has been performed on the efficiency and scalability of data mining methods. Moreover, the research has been extended to spatial data mining, multimedia data mining, financial mining, and Web mining with several new data mining system prototypes constructed or currently under construction, including GeoMiner, MultiMediaMiner, FinancialMiner, and WebLogMiner. This demo will show the most recent research and development status of the DBMiner system. Hopefully, the system will be available commercially by the time of demo.

What Is Unique about the System? On-line (interactive) analytical mining, multiple integrated data mining functions, integration of data mining with OLAP, and knowledge visualization tools.

Demos

Distributed Data Mining for Enterprise Solutions

Paper Title: An Enhanced KDD Process Model and its Visualisation

Development Team: M. Kolher, J. Chattratichat, Y. Guo and S. Hedvall, Imperial College

Tel.: +44-171-594-83-57

The Kensington System provides an enterprise solution for large-scale data mining in environments where data is logically and geographically distributed over multiple databases. Supported by an intuitive Integrated Programming/Visualisation Tool kit, an analyst explores remote databases and visually defines and executes procedures that model the entire KDD process. The system provides high performance components for the most common data mining tasks, such as classification, prediction, clustering, and association. Generated decision models are evaluated and modified using powerful interactive visualisation techniques.

Designed as a 3-tier application based on the Enterprise JavaBeans (EJB) architecture, application servers can be transparently distributed for scalability or replicated for increased availability. Defined KDD procedures and generated decision models are realized as persistent objects, which can easily be reused and shared between group members. Kensington imposes strong security on data transfer and model distribution through secure socket communications. Access control mechanisms protects user/group specific resources from unauthorized access.

For maximum flexibility and easy deployment, client tools are 100% Java compliant applets and runs securely in Web browsers everywhere on the\ Internet. A data analyst is therefore not bound to any specific location or computer.

What Is Unique about the System? (1) Novel programming/visualisation tools for Web clients; (2) scalable and flexible system architecture for Internet/Intranet environments; (3) secure data integration from distributed databases; (4) high performance modeling through parallelism.

Document Explorer and TextVis

Paper Titles: (1) TextVis: An Integrated Visual Environment for Text Mining *and* Text Mining at the Term Level*; (2)* Trend Graphs: Visualizing the Evolution of Concept Relationships in Large Document Collections

Development Team: Ronen Feldman, Yonatan Aumann, David Landau, Moshe Fresko, Orly Lipchtat, Yehuda Lindel, Yaron Ben Yehuda, Yonatan Schelr, Amir Zilberstein and Moshe Martziano, Bar-Ilan University and Instinct Software Ltd.

Tel.: +972-3-5318629

TextVis is a visual data mining system for document collections. Such a collection represents an application domain, and the primary goal of the system is to derive patterns that provide knowledge about this domain. Additionally, the derived patterns can be used to browse the collection. TextVis takes a multi-strategy approach to text mining, and enables defining complex analysis schemas from basic components, provided by the system. An analysis schema is constructed by dragging functional icons from a tool-pallette onto the workspace and connecting them according to the desired flow of information. The system provides a large collection of basic analysis tools, including: frequent sets, associations, concept distributions, and concept correlations. The discovered patterns are presented in a visual interface allowing the user to operate on the results, and to access the associated documents. TextVis is a complete text mining system which uses agent technology to access various online information sources, text preprocessing tools to extract relevant information from the documents, a variety of data mining algorithms, and a set of visual browsers to view the results.

What Is Unique about the System? A Unique collection of tools for Text Mining. A special set of Visual Maps. Easy Customization to match the exact needs of the user. Ability to build very complex Text analysis schemas.

DR — Data Mining Via Data Reduction

Paper Title: Data Reduction Based on Hyper Relations

Development Team: Hui Wang, University of Ulster

Tel.: +44-1232-368981

Data reduction makes datasets smaller but preserves classification structures of interest. In data mining, data reduction is regarded as a main task of data mining hence any data mining technique can be regarded as a method for data reduction (Usama Fayyad, 1997). We proposed a general (algebraic) approach to data reduction, which in turn can be used for data mining. A paper describing this approach is submitted to KDD-98. We have developed a system (called DR) based on this approach. We want to demonstrate DR with respect to the followings: (1) Data mining can be achieved via direct data reduction. (2) Data and models can be uniformly represented by hyper relations. (3) Datasets can be significantly reduced in size while the classification structures are preserved. (4) Attribute selection and discretization of continuous attributes can be achieved as a by-product of data reduction. (5) Missing values and overfitting can be naturally dealt with in DR. (6) DR can outperform C4.5 in many cases using public datasets

What Is Unique about the System? (1) Algebraic and theoretically well founded approach to data mining; (2) Uniform representation of data and models — hyper relations, a generalization of database relations in the traditional sense — therefore data mining can be taken to be an operation of database systems; (3) missing values and overfitting are naturally dealt with; (4) attribute selection and continuous attribute discretization can be achieved as a by-product; (5) the model built by DR can be further mined, if needed.

Finding Outliers with Depth Contours

Paper Title: Fast Computation of 2-Dimensional Depth Contours

Development Team: Raymond T. Ng and Ivy Kwok, University of British Columbia; Ted Johnson, AT&T Research Center

Tel.: 604-822-2394

"One person's noise is another person's signal." For many applications, including the detection of credit card frauds and the monitoring of criminal activities in electronic commerce, an important knowledge discovery problem is the detection of rare/exceptional/outlying events.

In computational statistics, one well-known approach to detect outlying data points in a 2-D dataset is to assign a depth to each data point. Based on the assigned depths, the data points are organized in layers in the 2-D space, with the expectation that shallow layers are more likely to contain outlying points than are the deep layers. One robust notion of depth, called depth contours, was introduced by Tukey [17,18]. ISODEPTH, developed by Ruts and Rousseeuw [16], is an algorithm that computes 2-D depth contours.

In this demo, we show a fast algorithm, called FDC, for computing 2-D depth contours. The idea is that to compute the first k depth contours, it is sufficient to restrict the computation to a small selected subset of data points, instead of examining all data points. Consequently, FDC scales up much better than ISODEPTH. For instance, for 1,000 data points FDC is 4 times faster than ISODEPTH, and for 5,000 points FDC is 50 times faster. While 100,000 points are too many for ISODEPTH to handle, FDC takes about 50 seconds to compute the first 20 depth contours.

Last but not least, ISODEPTH relies on the non-existence of collinear points. Removing all collinear points can be time consuming. FDC is robust against collinear points.

What Is Unique about the System? The last two paragraphs of the above description summarize the key points.

Frequent Substructure Discovery with WARMR

Paper Title: Finding Frequent Substructures in Chemical Compounds

Development Team: Luc Dehaspe, Hendrik Blockeel, Wim Van Laer and Luc De Raedt, Katholieke Universiteit Leuven; Hannu Toivonen, University of Helsinki

Tel.: +32-1632-7658

WARMR is a general purpose tool for the discov-

Demos

ery of frequent patterns, association rules at its simplest and first-order logic rules in the general case. We will demonstrate how both new and known variants of frequent pattern discovery are handled, and how the user can switch from one setting to another with minor efforts.

As a prototypical example of applications where the additional expressivity offered by WARMR is useful, we consider the discovery of frequent substructures in a biochemical database of compounds that are classified as being carcinogenic or not. In this context, patterns concern general properties of the compound, but also more complex features such as bonds between atoms, membership of atoms to chemical groups such as alcohols, and connections between chemical groups. Preparation of the experiment involves representation of the data and background knowledge in a DATALOG format, and the definition of the hypothesis space by means of a declarative language bias formalism.

Other applications with similar needs for highly expressive patterns are taken from the domains of (advanced) market basket analysis, discovery of linguistic knowledge in tree-banks, and telecommunication alarm analysis. WARMR is freely available for academic purposes upon request.

What Is Unique about the System? WARMR discovers useful frequent patterns that are way beyond the complexity of association rules or their known variants. By changing the language bias the user can easily search for different patterns without modifying the algorithm. The very natural facility to add background knowledge further enhances the flexibility of the tool.

InferView: Data Mining through Knowledge Inference and 3D Visualization

Development Team: Jerzy Bala, Mirco Manuci, Srinivas Gutta and Sung Baik, Datamat Systems Research, Inc.; Peter Pachowicz, George Mason University

Tel.: 703-917-0880, ext. 226

A research prototype of the InferView system will be demonstrated. InferView is being developed under the Ballistic Missile Organization and U.S. Army Space and Missile Defense Command sponsored project on data mining and decision support tools for situational awareness. InferView consists of three major components; (1) visualization, (2) inference engine, and (3) predictor. The user interacts with the system through a visual representation space where various graphical objects are rendered. Graphical objects represent: data, knowledge (e.g. as induced rules), and query explanations (decisions on unknown data identifications). InferView integrates graphical objects through the use of visually cognitive, human oriented depictions. A user can also examine non-graphical explanations (i. e. text based) to posed queries. InferView's transfer of data mining and decision support processes to the visualization space enhances the user's capabilities to see, explore, and gain decision making insights as never before.

The following two modes of operation will be demonstrated; (1) the data mining mode which uses the inference engine module to generate knowledge and subsequently represents it as 3D graphical objects, and (2) the decision support mode the predictor module is used to support user's queries. In both modes user's directed navigation, zooming, and other spatially oriented operations will be demonstrated.

What Is Unique about the System? The uniqueness of the InferView system is its synergistic integration of advanced computer graphic/visualization and inference based data generalization techniques. InferView's knowledge visualization techniques contribute to better human decision-making insights through facilitation of spatial operations such as navigation, zooming, etc. A graphically appealing human computer interfacing and capability to visualize large and complex knowledge bases through spatial and graphical depictions of knowledge components add to InferView's uniqueness.

Interactive Mining of Interesting Rules

Papers Titles: (1) Integration of Classification and Association Rule Mining; (2) Visual Aided Exploration of Interesting Association Rules

Development Team: Bing Liu, Wynne Hsu, Yiming Ma and Chen Shu, National University of Singapore

Tel.: +65-874-6736

We would like to demonstrate two main themes of our data mining system: (1) Using association rule mining technique to find all potential classification rules (PCRs), and then building an accurate classifier using the PCRs. This shows that classification and association rule mining can be integrated. In an application, the user can use one system for two purposes. (2) Helping the user to identify subjectively interesting rules. The number of association rules or potential classification rules that exist in a database can be huge. This makes manual inspection and analysis of the rules difficult. We have designed and implemented a new framework to allow the user to visually explore the discovered rules to identify those interesting ones easily. This framework has two components, an interestingness analysis component, and a visualization component. The interestingness analysis component analyzes and organizes the discovered rules according to various interestingness criteria with respect to the user's existing knowledge. The visualization component enables the user to visually explore those potentially interesting rules. Enhanced with color effects, the user can easily and quickly focus his/she attention on the more interesting/useful rules.

What Is Unique about the System? (1) Our system integrates association rule mining and classification rule mining. We have adapted an association rule technique to discover all potential classification rules. We then use this complete set of rules to build an accurate classifier; (2) We have designed and implemented a post-analysis system to help the user identify subjectively interesting association rules and classification rules with the aid of a visualization component.

Telecommunications Churn Analysis Using the Amdocs KDD Environment

Paper Title: Ranking - Methods for Flexible Evaluation and Efficient Comparison of

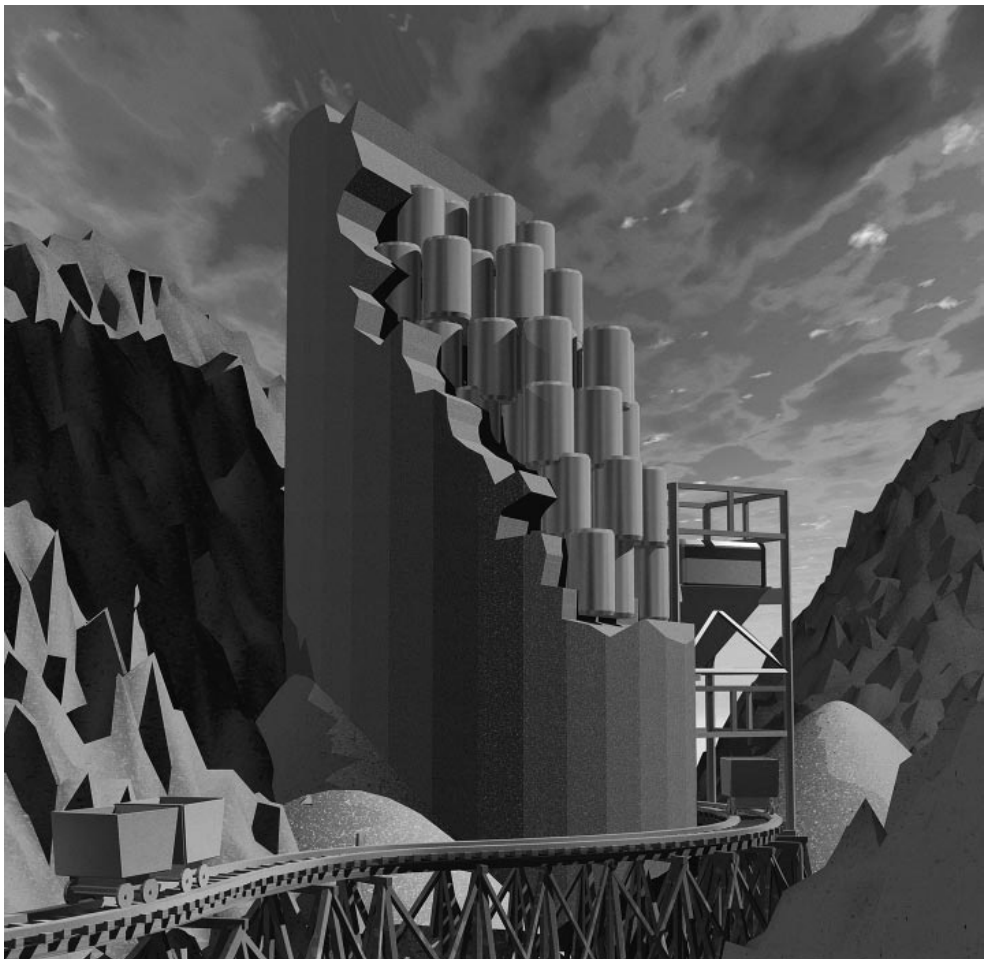
Classification Performance

Development Team: Dr. Gadi Pinkas, Dr. Yizhak Idan, Rony Paz and Saharon Rosset, Amdocs Inc.

Tel.: +972-3-5765174

The demonstration will present a full modeling and analysis process of churn data coming from a telecommunication operator data warehouse. The demonstration will include: (1) Extraction and reformulation of relevant data from the customer's Data Warehouse, including formation of time-dependent and dynamic indicators as new input fields; (2) Use of different definitions for churn (percentage of lines, fall in usage, etc.); (3) Insight into man machine interface for the analysis of the Data-Mining results and the incorporation of business knowledge. In this stage the human analyst/expert can examine the rules which the system generated, modify them and add other rules expressing his prior/external knowledge; (4) Model construction and visualization tools that include multiple model-building algorithms, based on the combination of automatically discovered and handcrafted rules. These innovative algorithms combine machine learning (Rule Induction) and statistical (Logistic Regression) methods to provide a hybrid classifier experimentally proven to be highly effective; (5) Emphasis on the customer's value (rather than just his chances of performing churn) both in constructing the classifier and in evaluating it. The use of value is integrated into the automated discovery process, so the rules generated actually predict revenue flow rather than just customer flow.

What Is Unique about the System? (1) Visualization and analysis module for combining the automated discovery results with human expert knowledge; (2) Optimization with regard to revenue flow rather than customer flow. This is in contrast with actual methods limitations that either introduce value into post-DM analysis or perform pre-DM segmentation by value.



ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING

Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy editors

ISBN 0-262-56097-6 632 pp., index. $50.00 softcover

The AAAI Press • Distributed by The MIT Press
Massachusetts Institute of Technology, 5 Cambridge Center Cambridge, Massachusetts 02142
To order, call toll free: (800) 356-0343 or (617) 625-8569.
MasterCard and VISA accepted.