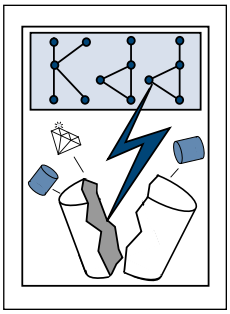

The Second International Conference on Knowledge Discovery and Data Mining—KDD-96

Product Demonstrations

Ac2: Advanced Decision Tree-based Data Mining

Cyril Way, Hugues Marty, Thierry Marie Victoire ISoft

Ac2 is an interactive decision tree based data mining development environment with advanced knowledge modeling capabilities. It includes an object oriented graphical language for structuring raw data and guiding the data mining process with domain knowledge. Data being structured or not, the tool builds decision trees automatically. Users may freely interact with information displayed in the trees, by modifying tests, cutting or expanding tree branches, grouping or ungrouping tree nodes. Reporting features include graphs and text reports. Various ID3 and CART based algorithms are available within the tool.

Unique features: OO graphical language for knowledge modeling. No software limitations on data volumes to be processed. Fully interactive development tool and C++ libraries.

Status: Commercial product.

Clementine Data Mining System

Colin Shearer Data Mining Division, Integral Solutions Ltd.

Launched in 1994, Clementine is a comprehensive end-user toolkit for data mining. It integrates multiple modeling technologies—neural networks, rule induc- tion, and statistical regression—with interactive visualization modules, database/file access and numerous data pre-processing/manipulation facilities. Clementine is driven using a sophisticated visual programming interface; the user selects icons representing data sources and operations, and positions and connects these to specify data flows. The modeling modules are by default automatically configured, based on the data presented. Expert options allow fine control of the algorithms, and an intermediate level allows the user to specify high-level strategies or preferences. Models built within a Clementine session can be exported as C source code for embedding within other applications. Clementine is in use in numerous areas, with applications in finance, retail, energy, science/pharmaceuticals, marketing, manufacturing, government, telecommunications and defense.

Unique features: Integration of many algorithms/facilities. Accessibility to non-technologist end-users through visual programming and automatic configuration.

Status: Commercial product.

DataMine—An Integrated Geomarketing Decision Support System

Cyril Way, Hugues Marty, Thierry Marie Victoire, ISoft

Geomarketing decision support system in a banking organization. CAISSE D'EPARGNE, one of the top three banks in France, operates a 3,000 automated teller machines network. These ATMs may generate profits, or losses: the better the location, the higher the return on investment and the lower the risk. The application built by ISoft is a global data mining application, that aims at providing the bank's managers with forecasts of new ATMs profitability, in order to help them decide whether the machine should be set up or not. The development is made of several steps:

Knowledge modeling: Which parameters may explain an ATM's behavior? Data collection and preparation: gathering data and controlling their integrity.

Data mining: Extracting models from raw data, testing and validating them.

Implementing these models: making them available to every manager through a distributed decision support application.

Unique features: Auto improvement of the prediction accuracy rate. Little obsolescence risk.

Status: Fielded application.

Data Surveyor

M.Holsheimer, F.Kwakkel, D.Kwakkel, P. Boncs. Data Distilleries, Kruislaan 419, 1098 VA Amsterdam, The Netherlands.

Data Surveyor is a data mining tool that has been developed by Data Distilleries in close co-operation with CWI in the Netherlands and other members of the EC-funded KESO (knowledge engineering for statistical offices) research project. We demonstrate the analytical, graphical and reporting capabilities of Data Surveyor by the interactive analysis of a number of datasets, constructed on the basis of our experience with clients. Example applications include the discovery of risk profiles in an insurance data base, and in a credit granting data base. We demonstrate the possibility to mine interactively on large databases, allowing the user to further explore interesting elements of the discovered knowledge.

Unique features: Data Surveyor allows the user to mine interactively on large data bases, by using efficient search strategies and database optimization techniques. Profiles generated from the database are easily understood by the user, thus enhancing the interactive and creative character of the data mining exercise. Visualization techniques are used to further enhance the user's insight.

Status: Commercial product.

DBMiner: A System for Mining Knowledge in Large Relational Databases

Jiawei Han, Yongjian Fu, Wei Wang, Jenny Chiang, Wan Gong, Krzysztof Koperski, Deyi Li, Yijun Lu, Amynmohamed Rajan, Nebojsa Stefanovic, Betty Xia, Osmar R. Zaiane.School of Computing Science Simon Fraser University, British Columbia, Canada V5A 1S6

A data mining system, DBMiner, has been developed for interactive mining of multiple-level knowledge in large relational databases. The system implements a wide spectrum of data mining functions, including generalization, characterization, association, classification, and prediction. By incorporation of several interesting data mining techniques, including attribute-oriented induction, statistical analysis, progressive deepening for mining multiple-level knowledge, and meta-rule guided mining, the system provides a user-friendly, interactive data mining environment with good performance.

Unique features: It implements a wide spectrum of data mining functions including generalization, characterization, association, classification, and prediction. It performs interactive data mining at multiple concept levels on any user-specified set of data in a database using an SQL-like data mining query language, DMQL, or a graphical user interface, and visual presentation of rules, charts, curves, etc. Both UNIX and PC (Windows/NT) versions of the system adopt a client/server architecture.

Status: Research prototype.

Decisionhouse

Quadstone Ltd.

Decisionhouse is an integrated software suite for scalable data analysis and visualization. It offers integration with the major RDBMS packages, fast segmentation and profiling, geodemographic analysis, pre- and post-processing and interactive 3-dimensional visualization. Although the functionality provided is generic, Decisionhouse is currently principally targeted as a tool for analyzing large customer databases.

Unique features: Parallelism, speed, integration.

Status: Commercial product.

D-SIDE: A Probabilistic Decision Endorsement Environment

P.Kontkanen, P.Myllym?ki and H.Tirri: Complex Systems Computation Group Department of Computer Science University of Helsinki, Finland

D-SIDE (a probabilistic decision endorsement environment) is a generic tool for probabilistic inference and model construction. The computational model is based on finite mixture distributions, which are constructed by clustering data into groups of similar elements. The D-SIDE software is a platform independent prototype with a graphical JAVA user interface, which allows the software to be used across the Internet. D-SIDE offers an elegant solution for both prediction and data mining problems. For prediction purposes, D-SIDE provides a computationally efficient scheme for computing Bayes optimal predictive distributions for any set of attributes. In data mining applications, the user can explore attribute dependencies by visualizing the cluster structure found in data. In empirical tests with various public domain classification datasets, D-SIDE consistently outperforms the results obtained by alternative approaches, such as decision trees and neural networks. A running demo of the D-SIDE software is available for experimentation through our WWW homepage at URL http:// www.cs.Helsinki.FI/research/cosco.

Unique features: The system combines the computationally efficient, yet expressive set of finite mixture models with a theoretically solid Bayesian approach for model construction and prediction. The interface is based on the platform independent JAVA language, which allows the software to be used across the Internet.

Status: Research prototype.

FACT : Finding Associations in Collections of Text

Ronen Feldman, Bar-Ilan University, and Haym Hirsh, Rutgers University.

FACT (finding associations in collections of text) is a system for knowledge discovery from text. It discovers associations— patterns of co-occurrence—among keywords labeling the items in a collection of textual documents. In addition, when background knowledge is available about the keywords labeling the documents FACT is able to use this information in its discovery process FACT takes a query-centered view of knowledge discovery, in which a discovery request is viewed as a query over the implicit set of possible results supported by a collection of documents, and where background knowledge is used to specify constraints on the desired results of this query process. Execution of a knowledge-discovery query is structured so that these background-knowledge constraints can be exploited in the search for possible results. Finally, rather than requiring a user to specify an explicit query expression in the knowledge-discovery query language, FACT presents the user with a simple-to-use graphical interface to the query language, with the language providing a well-defined semantics for the discovery actions performed by a user through the interface.

Status: Research prototype.

IBM Data Mining Tools

Julio Ortega, Kamal Ali, Stefanos Manganaris, George John, IBM Almaden Research Center

IBM has developed several powerful algorithms and processing techniques that enable application developers to analyze data stored in databases or flat files. These algorithms enable analyses ranging from classification and predictive modeling, to association discovery, and database segmentation. Using predictive modeling, a retailer could forecast changes in customer buying patterns or characterize different types of customers, for example, those that are likely to switch from in-store buying to internet or mail-order buying. Through association discovery, a supermarket chain could determine which products are most frequently sold in conjunction with other products. More details about can be found at http://www.almaden.ibm. com/stss/.

Unique features: The IBM Data Mining tools provide access to a comprehensive set of data manipulation functions and machine learning algorithms, thus allowing developers and consultants to quickly build customized data mining applications. IBM's Data Mining tools can run on small workstations but is they are highly scalable since the algorithms have parallel implementations optimized to handle large and parallel databases.

Status: Commercial product.s.

Kepler: Extensibility in Data Mining Systems

Stefan Wrobel, Dietrich Wettschereck, Edgar Sommer,
Werner Emde GMD, FIT.KI, Schloss Birlinghoven, 53754
Sankt Augustin, Germany

Given the variety of tasks and goals of data mining applications, there can never be a fixed arsenal of analysis methods suffi-

cient for all applications. Many applications even require custom methods to be used. Systems with a fixed selection of methods, or maybe just one single method, are thus likely to be outgrown quickly, forcing the analyst to switch to a different system, incurring time-consuming training and data conversion overhead. In our demonstration, we show Kepler, an integrated data mining system extensible through a "plug-in" facility. The plug-in facility allows algorithms for existing or new analysis tasks to be rapidly plugged into the system without redeveloping the system core.

Already available are plug-in algorithms for decision tree induction, backprop neural networks, subgroup discovery, clustering, nearest neighbor learning, and multiple adaptive regression splines. We illustrate the system in the context of our applications on retail data analysis and ecological modeling.

Unique features: Kepler is easily extensible with new methods, new analysis tasks, new preprocessing operators through its plug-in facility and already offers a wide variety of tasks and methods.

Status: Research prototype.

Knowledge Discovery Environment

Aashu Virmani, Amin Abdulghani and Tomasz Imielinski).Department of Computer Science, Rutgers University

We will demonstrate the process of querying a rulebase, (rule-mining) and then using the same query to generate that set of rules from the database (data-mining). The process of "mining around a rule," where the answer from a query can be resubmitted as a query to explore deeper into the rule, will also be demonstrated. Next, we will show some sample applications like "typicality," "distinguishing characteristics of" and "best classifiers for" which are built on top of rules, proving that rule-generation is not the final step in data-mining. We will then show the "exploratory mining scenario," in which the user can query the rulebase as it is being generated, narrow his preferences about what he wants, and prune the overall rule-space the system must generate. Finally, we shall try to unify all the above concepts in a programming environment, which makes use of a C++ API, and allows KDD-developers and users alike, to embed rule generation in larger applications, and also provides support for defining new attributes and incorporate them in the mining queries. We will code a few quick applications and run them as part of the demo.

Unique features: It allows an ad-hoc query based approach to mining, where the query can be used both for rule selection, and rule generation. It demonstrates some higher level applications that can be developed when we treat rules as first class objects. It provides a C++ API, and a programming environment, which allows users/developers to build complex kdd-applications.

Status: Research prototype.

Management Discovery Tool

Ken O'Flaherty, NCR Corporation , Parallel Systems Division

The NCR Management Discovery Tool (MDT) is a new class of business reporting tool which applies intelligence not present in today's tools to provide actionable information for managers and other users of the data warehouse. MDT combines three advanced technologies to create management reports: business knowledge representation, dynamic user interfaces and database query generation. These reports summarize situations, trends and events occurring in your business. Working from customized business rules defined by your company's knowledge workers, MDT highlights and explains changes and trends for your key business indicators. MDT also monitors selected business indicators for deviations from plan, and can alert you to exceptions as they occur, with automatic analysis of the causes driving the exceptions. Information is generated and presented in the form of InfoFrames. These presentation-quality compound documents consist of natural language reports and explanatory graphics that are dynamically generated by MDT.

Unique features: MDT is unique in its ability to automatically retrieve relevant information of importance to one's business, and incorporate it dynamically into concise high-quality management reports. This information is in three categories: drill-across—to surrounding segments of interest, drill-down—to selected segments that make up the target segment and causal analysis—indicating probable causes of the business situation.

Status: MDT is currently in development, with a planned release in the fourth quarter of 1996. NCR will demonstrate a prototype version.

MineSet

Steven Reiss and Mario Schkolnick, Data Mining and Visualization Group, Silicon Graphics Computer Systems

MineSet is a suite of tools that allows the user to access data from data warehouses, transform the data, apply data mining algorithms to the data, and then use visualization to perform visual data mining. It supports direct access to commercial RDBMS databases. MineSet provides the user with ability to transform data through the use of binning, aggregations, and groupings. MineSet supports the generation of both classification (decision tree and naive Bayes) and association rule models directly from the data.

Four visualization tools are provided with MineSet. These visual data mining tools enable analysts to interactively explore data and quickly discover meaningful new patterns, trends, and relationships. The tools support both spatial and non-spatial trend analysis through animation, discovery of clustering and profiling, analysis of hierarchical data using 3D flythrough, and visualization and analysis of data mining results. The MineSet Tool Manager serves as the command console for the MineSet user. Using the Tool Manager, one can specify what data to access, how to transform the data, how to mine the data, and how to visualize the results of the data transformation and mining process.

Unique features: MineSet is unique in its integration of data mining and visualization. Visualization serves two purposes. It provides the user with the ability to do visual data mining, allowing discovery of trends, patterns, and anomalies contained in vast amounts of data. MineSet also allows for better understanding, analysis, and exploration of classification and association models resulting from the data mining process.

Status: Commercial product.

MM — Mining with Maps

Raymond Ng Computer Science, University of British Columbia

This package features three spatial data mining algorithms. The first algorithm computes aggregate proximity relationships. The second algorithm discovers characteristic and discriminating properties of spatial clusters from thematic maps. The third algorithm performs boundary shape matching in spatial data. The package also includes visualization tools. More details of these algorithms

can be found in "Finding Aggregate Proximity Relationships and Commonalties in Spatial Data Mining" to appear in: *IEEE Transactions on Knowledge and Data Engineering, Special Issue on Data Mining;* "Extraction of Spatial Proximity Patterns by Concept Generalization" to appear in the *KDD-96 Proceedings;* and "Spatial Data Mining: Discovering Knowledge of Clusters from Maps" to appear in: *Proceedings, 1996 SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery.*

Unique features: All three spatial data mining algorithms are unique. The package is, therefore, the first implementation of the three algorithms.

Status: Research prototype.

Optimization Related Data Mining Using the PATTERN System

R. L. Grossman, H. Bodek, D. Northcutt Magnify, Inc.and H. V. Poor, Princeton University

We will demonstrate a system which integrates an object warehouse with data mining and data analysis tools. The system is designed to work with large volumes of data. We have developed an object warehouse which is specifically designed for the types of queries and analyses arising in data mining. Because of the amount of data analyzed, rather than search for any patterns, the system narrows the search to patterns of interest in optimization problems. We will demonstrate the system on several applications, including anomaly detection, mining textural data, and credit card account management.

Unique features: We integrated an object warehouse with data mining tools. Using an object warehouse and specialized algorithms we can mine large data sets. We have developed and implemented distributed and parallel versions of tree-based data mining algorithms. We have developed specialized algorithms for finding patterns related optimization.

Status: Commercial product, currently available in beta with several customers

STARC

Damir Gainanow, Andre Matweew, Michael Thess DATA-Center Ltd., Ekaterinburg, RUSSIA Scholz & Thess Software GbR, Chemnitz, GERMANY

Rapid developments in the field of science and technology in recent years have created an enormous amount of electronic data. The accompanying improvements in the generation, transport and storage of this data have opened up vast opportunities demanding new and more efficient data processing methods. This calls for improved tools which, with the aid of data mining methods, are capable of identifying novel, potentially useful, and ultimately understandable patterns in existing data. Such a knowledge discovery system has now been developed under the name of STARC (statistics, analysis, recognition, clustering) by the Russian company DATA-CENTER headed by D. Gainanow. STARC offers the user a wide variety of modules which can be combined in numerous ways to create a powerful data processing tool. Different modules provide different methods of statistics, supervised and unsupervised machine learning, machine discovery, data compression and data visualization.

Unique features: Most existing decision tree methods are based on the minimization of the number of misclassified vectors in the feature space. In this context, there is the problem of getting stuck in only local minima. To overcome this shortcoming an outstanding novel algebraic approach is applied in the system STARC.

Status: Commercial product.

WebFind: Mining External Sources to Guide WWW Discovery

Alvaro E. Monge and Charles P. Elkan Department of Computer Science and Engineering University of California, San Diego La Jolla, CA 92093-0114 amonge,elkan@cs.ucsd.edu

WebFind is an application that discovers scientific papers made available by their authors on the worldwide web. WebFind mines external sources for information that aids in the discovery process. Currently there are two information sources being mined: Melvyl and Netfind. Melvyl is a University of California library service that includes a comprehensive database of bibliographic records. Netfind is a white pages service that gives internet host addresses and email address. Separately these services do not provide enough information to locate papers on the worldwide web. WebFind integrates the information provided by each in order to discovery on the worldwide web the information actually wanted by a user. A WebFind search starts with the user providing keywords identify the paper, exactly as he or she would in searching Inspec directly. After the user confirms that the right paper has been identified, WebFind queries Inspec to find the institutional affiliation of the principal author of the paper. Then WebFind uses Netfind to provide the Internet address of a host computer with the same institutional affiliation. WebFind then uses a search algorithm to discover a worldwide web server on this host, then an author's home page, and finally the location of the wanted paper. Since institutions are designated very differently in Inspec and Netfind, it is nontrivial to decide when an Inspec institution corresponds to a Netfind institution. WebFind uses a recursive field matching algorithm to do this.

Unique features: Automatic discovery of information on the worldwide web is achieved in real-time. Unlike typical search engines, WebFind does not use a central index of the worldwide web. It is able to do this through mining of information from other sources. At the heart of this mining is the field matching algorithm, which identifies semantically equivalent information in multiple sources.

Status: Research Prototype accessible through the WWW at: http://dino.ucsd.edu:8000/WebFind.html

WEBSOM—Interactive Exploration of Document Collections

Krista Lagus, Timo Honkela, Samuel Kaski and Teuvo Kohonen, Neural Networks Research Centre of Helsinki University of Technology

WEBSOM is a new approach to exploring collections of full-text documents. A map of a document collection where similar documents are found close to each other can be explored with an easy interface. The demonstration is also accessible to anyone in the address http://websom.hut.fi/websom/.

Unique features: Visualization of the document space with a map where related documents appear close to each other. The method by which such a meaningful map can be produced.

Status: Research prototype.