A SURVEY OF DATA MINING AND KNOWLEDGE DISCOVERY SOFTWARE TOOLS

Michael Goebel

Department of Computer Science
University of Auckland
Private Bag 92019, Auckland
New Zealand

mgoebel@cs.auckland.ac.nz

Le Gruenwald

School of Computer Science
University of Oklahoma
200 Felgar Street, Room 114
Norman, OK, 73019

gruenwal@cs.ou.edu

ABSTRACT

Knowledge discovery in databases is a rapidly growing field, whose development is driven by strong research interests as well as urgent practical, social, and economical needs. While the last few years knowledge discovery tools have been used mainly in research environments, sophisticated software products are now rapidly emerging. In this paper, we provide an overview of common knowledge discovery tasks and approaches to solve these tasks. We propose a feature classification scheme that can be used to study knowledge and data mining software. This scheme is based on the software's general characteristics, database connectivity, and data mining characteristics. We then apply our feature classification scheme to investigate 43 software products, which are either research prototypes or commercially available. Finally, we specify features that we consider important for knowledge discovery software to possess in order to accommodate its users effectively, as well as issues that are either not addressed or insufficiently solved yet.

Keywords

Knowledge discovery in databases, data mining, surveys.

1. INTRODUCTION

1.1 Overview and Motivation

The rapid emergence of electronic data management methods has lead some to call recent times as the "Information Age." Powerful database systems for collecting and managing are in use in virtually all large and mid-range companies -- there is hardly a transaction that does not generate a computer record somewhere.

Each year more operations are being computerized, all accumulate data on operations, activities and performance. All these data hold valuable information, e.g., trends and patterns, which could be used to improve business decisions and optimize success.

However, today's databases contain so much data that it becomes almost impossible to manually analyze them for valuable decision-making information. In many cases, hundreds of independent attributes need to be simultaneously considered in order to accurately model system behavior. Therefore, humans need assistance in their analysis capacity.

This need for automated extraction of useful knowledge from huge amounts of data is widely recognized now, and leads to a rapidly developing market of automated analysis and discovery tools. Knowledge discovery and data mining are techniques to discover strategic information hidden in very large databases. Automated discovery tools have the capability to analyze the raw data and present the extracted high level information to the analyst or decision-maker, rather than having the analyst find it for himself or herself.

In the last few years, knowledge discovery and data mining tools have been used mainly in experimental and research environments. Now we are at a stage where sophisticated tools, which aim at the mainstream business user, are rapidly emerging. New tools hit the market nearly every month. The Meta Group estimates that the market size for data mining market will grow from $50 million in 1996 to $800 million by 2000 ([35]).

The aim of this study is fourfold:

1) Provide an overview of existing techniques that can be used for extracting of useful information from databases.
2) Provide a feature classification scheme that identifies important features to study knowledge discovery and data mining software tools.
3) Investigate existing knowledge discovery and data mining software tools using the proposed feature classification scheme. These tools may be either commercial packages available for purchasing, or research prototypes developed at various universities.
4) Identify the features that discovery software should possess in order to accommodate novice users as well as experienced analysts.

The rest of this paper is organized as follows. Section 1.2 lists categories of discovery software that are intentionally not included in this report. Section 1.3 lists some other currently ongoing survey projects and available information resources. Section 2 discusses the process of analyzing high-volume data and turning them into valuable decision support information. It describes the various analysis tasks as well as the techniques for solving these tasks. Section 3 presents the proposed feature classification scheme and the review of existing software using this scheme (the detailed descriptions for each product are provided in [24]). Section 4 draws some conclusions about the current state of existing discovery tools, and identifies some desirable features and characteristics that make a discovery tool truly useful, thus providing directions for future research.

1.2 Software not included in this paper

A major goal of our study is to provide a market overview of off-the-shelf software packages whose main purposes are to aid in knowledge discovery and data mining. In order to keep the scope

of this paper focused, the following items are intentionally not considered in this study:

- Software that exclusively acts as information server to data mining tools and does not perform analysis itself, such as Geneva from Price Waterhouse LLP ([43]).
- Software that can be "abused" for data mining, but its intended use lies somewhere else, such as MATLAB's Neural Network Toolbox ([16]) or statistical software packages.
- Software that is marketed as data mining or information discovery software, but in reality is not much more than a reporting or visualization tool, such as the Oracle Discoverer ([41]).
- Many Companies offer consulting services and development of industry-specific solutions. This survey however is only about off-the-shelf products.

While we tried to make the information in this survey as complete and accurate as possible, we also found quite some companies that did not want to disclose their technologies and algorithms used because of competitive advantages.

1.3 Other Survey Projects

Some attempts to provide surveys of data mining tools have been made, for example:

- The Data Mine ([45]) includes pointers to downloadable papers, and two large data mining bibliographies. It attempts to provide links to as much of the available data mining information on the net as is possible.
- The Knowledge Discovery Mine ([44]) has the KDD FAQ, a comprehensive catalog of tools for discovery in data, as well as back issues of the KDD-Nuggets mailing list.
- The Exclusive Ore internet site ([18]) contains a data mining product features table that compares key features of approximately 15 of the top products and has links directly from the features table to each product. The site, which is still under construction, will also have tutorials on various data mining technologies and problems, illustrated with real examples that, at the same time, show how various products work.
- Kurt Thearling maintains a set of knowledge discovery related WWW pages ([56]). Among others, there is a list more than 60 data mining software vendors, a list with software patents related to data mining, and general information (tutorials and papers) related to data mining.

In our work, we want to provide a method to study software tools and apply this method to investigate a comprehensive set of 43 existing tools. For each tool, we examine key features and provide detailed descriptions as well as summary tables. Our aim is to facilitate software developers as well as prospective data mining users.

2. KNOWLEDGE DISCOVERY AND DATA MINING

This section provides an introduction into the area of knowledge discovery, and serves as background explanation for our feature classification scheme and the software feature tables in section 3.

After we briefly elucidate our use of the terms knowledge discovery and data mining in section 2.1, we examine issues related to the database connection of the discovery endeavor in section 2.2. In section 2.3 we list the various analysis tasks that can be goals of a discovery process. Finally, section 2.4 lists methods and research areas that are promising in solving these analysis tasks.

2.1 The Knowledge Discovery Process

There is still some confusion about the terms *Knowledge Discovery in Databases (KDD)* and *data mining*. Often these two terms are used interchangeably. We use the term *KDD* to denote the overall process of turning low-level data into high-level knowledge. A simple definition of KDD is as follows: *Knowledge discovery in databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data* ([20]). We also adopt the commonly used definition of data mining as the extraction of patterns or models from observed data. Although at the core of the knowledge discovery process, this step usually takes only a small part (estimated at 15% to 25 %) of the overall effort ([8]). Hence data mining is just one step in the overall KDD process. Other steps for example involve:

- Developing an understanding of the application domain and the goals of the data mining process
- Acquiring or selecting a target data set
- Integrating and checking the data set
- Data cleaning, preprocessing, and transformation
- Model development and hypothesis building
- Choosing suitable data mining algorithms
- Result interpretation and visualization
- Result testing and verification
- Using and maintaining the discovered knowledge

There are several very good papers on the overall endeavor of knowledge discovery. The interested reader may refer to ([8], [19], [20], or [21]) as good introductory readings.

2.2 Database issues

Any realistic knowledge discovery process is not linear, but rather iterative and interactive. Any one step may result in changes in earlier steps, thus producing a variety of feedback loops. This motivates the development of tools that support the entire KDD process, rather than just the core data-mining step. Such tools require a tight integration with database systems or data warehouses for data selection, preprocessing, integrating, transformation etc.

Many tools currently available are generic tools from the AI or statistics' community. Such tools usually operate separately from the data source, requiring a significant amount of time spent with data export and import, pre- and post-processing, and data transformation. However, a tight connection between the knowledge discovery tool and the analyzed database, utilizing the existing DBMS support, is clearly desirable.

For the reviewed knowledge discovery tools, the following features are inspected:

Ability to access a variety of data sources: In many cases, the data to be analyzed is scattered throughout the corporation, it has to be gathered, checked, and integrated before a meaningful analysis

can take place. The capability to directly access different data sources can thus greatly reduce the amount of data transforming.

Online/Offline data access: Online data access means that queries are run directly against the database and may run concurrently with other transactions. In offline data access the analysis is performed with a snapshot of the data source, in many cases involving an export/import process from the original data source to a data format required by the discovery tools. The question of online or offline data access becomes especially important when one has to deal with changing knowledge and data: In financial markets for example, rapidly changing market conditions may make previously discovered rules and patterns invalid.

The underlying data model: Many tools that are available today just take their input in form of one table, where each sample case (record) has a fixed number of attributes. Other tools are based on the relational model and allow querying of the underlying database. Object-oriented and nonstandard data models, such as multimedia, spatial or temporal, are largely beyond the scope of current KDD technology ([21]).

Maximum number of tables/rows/attributes: These are theoretical limits on the processing capabilities of the discovery tool.

Database size the tool can comfortably handle: The anticipated amount of data to be analyzed should be an important factor in choosing a discovery tool. While the maximum numbers of tables/rows/attributes are theoretical limitations, there are also practical limitations that are posed by computing time, memory requirements, expressing and visualization capabilities etc. A tool that holds all data in main memory for example may be not appropriate for very large data sources, even if the theoretical maximum number of rows is unlimited.

Attribute types the tool can handle: Some discovery tools have restrictions upon the attribute types of the input data. For example, tools based on neural networks usually require all attributes to be of numeric type ([33]). Other approaches such as [49] may not be able to handle continuous (real) data, etc. Therefore, the attribute types present in the data source should be considered when selecting an analysis tool.

Query language: The query language acts as an interface between the user and the knowledge and database. It allows the user to process data and knowledge and to direct the discovery process. Some tools do not have a query language: human interaction is restricted to the specification of some process parameters. Others allow querying of the data and/or knowledge via queries formulated in some query language, which may be a standard language like SQL or an application specific language. Querying of data and knowledge may also take place via a graphical user interface (GUI).

The reviewed tools differ greatly in each of the features described above. The choice of a tool therefore depends on application specific requirements and considerations, such as form and size of the data available, goals of the discovery process, needs and training of the end user, etc.

2.3 Data Mining Tasks

At the core of the KDD process are the data mining methods for extracting patterns from data. These methods can have different goals, dependent on the intended outcome of the overall KDD process. It should also be noted that several methods with different goals may be applied successively to achieve a desired result. For example, to determine which customers are likely to buy a new product, a business analyst might need to first use clustering to segment the customer database, then apply regression to predict buying behavior for each cluster.

Most data mining goals fall under the following categories:

Data Processing: Depending on the goals and requirements of the KDD process, analysts may select, filter, aggregate, sample, clean and/or transform data. Automating some of the most typical data processing tasks and integrating them seamlessly into the overall process may eliminate or at least greatly reduce the need for programming specialized routines and for data export/import, thus improving the analyst's productivity.

Prediction: Given a data item and a predictive model, predict the value for a specific attribute of the data item. For example, given a predictive model of credit card transactions, predict the likelihood that a specific transaction is fraudulent. Prediction may also be used to validate a discovered hypothesis.

Regression: Given a set of data items, regression is the analysis of the dependency of some attribute values upon the values of other attributes in the same item, and the automatic production of a model that can predict these attribute values for new records. For example, given a data set of credit card transactions, build a model that can predict the likelihood of fraudulence for new transactions.

Classification: Given a set of predefined categorical classes, determine to which of these classes a specific data item belongs. For example, given classes of patients that correspond to medical treatment responses, identify the form of treatment to which a new patient is most likely to respond.

Clustering: Given a set of data items, partition this set into a set of classes such that items with similar characteristics are grouped together. Clustering is best used for finding groups of items that are similar. For example, given a data set of customers, identify subgroups of customers that have a similar buying behavior.

Link Analysis (Associations): Given a set of data items, identify relationships between attributes and items such as the presence of one pattern implies the presence of another pattern. These relations may be associations between attributes within the same data item (*'Out of the shoppers who bought milk, 64% also purchased bread'*) or associations between different data items (*'Every time a certain stock drops 5%, a certain other stock raises 13% between 2 and 6 weeks later'*). The investigation of relationships between items over a period of time is also often referred to as 'sequential pattern analysis'.

Model Visualization: Visualization plays an important role in making the discovered knowledge understandable and interpretable by humans. Besides, the human eye-brain system itself still remains the best pattern-recognition device known. Visualization techniques may range from simple scatter plots and histogram plots over parallel coordinates to 3D movies.

Exploratory Data Analysis (EDA): Exploratory data analysis (EDA) is the interactive exploration of a data set without heavy dependence on preconceived assumptions and models, thus attempting to identify interesting patterns. Graphic representations of the data are used very often to exploit the power of the eye and human intuition. While there are dozens of software packets

available that were developed exclusively to support data exploration, it might also be desirable to integrate these approaches into an overall KDD environment.

2.4 Data Mining Methodology

It should be clear from the above that data mining is not a single technique, any method that will help to get more information out of data is useful. Different methods serve different purposes, each method offering its own advantages and disadvantages. However, most methods commonly used for data mining can be classified into the following groups.

Statistical Methods: Historically, statistical work has focused mainly on testing of preconceived hypotheses and on fitting models to data. Statistical approaches usually rely on an explicit underlying probability model. In addition, it is generally assumed that these methods will be used by statisticians, and hence human intervention is required for the generation of candidate hypotheses and models.

Case-Based Reasoning: Case-based reasoning (CBR) is a technology that tries to solve a given problem by making direct use of past experiences and solutions. A case is usually a specific problem that has been previously encountered and solved. Given a particular new problem, case-based reasoning examines the set of stored cases and finds similar ones. If similar cases exist, their solution is applied to the new problem, and the problem is added to the case base for future reference.

Neural Networks: Neural networks (NN) are a class of systems modeled after the human brain. As the human brain consists of millions of neurons that are interconnected by synapses, neural networks are formed from large numbers of simulated neurons, connected to each other in a manner similar to brain neurons. Like in the human brain, the strength of neuron interconnections may change (or be changed by the learning algorithm) in response to a presented stimulus or an obtained output, which enables the network to "learn".

Decision Trees: A decision tree is a tree where each non-terminal node represents a test or decision on the considered data item. Depending on the outcome of the test, one chooses a certain branch. To classify a particular data item, we start at the root node and follow the assertions down until we reach a terminal node (or leaf). When a terminal node is reached, a decision is made. Decision trees can also be interpreted as a special form of a rule set, characterized by their hierarchical organization of rules.

Rule Induction: Rules state a statistical correlation between the occurrence of certain attributes in a data item, or between certain data items in a data set. The general form of an association rule is *Xl ^ ... ^ Xn => Y [C, S]*, meaning that the attributes *Xl,...,Xn* predict *Y* with a confidence *C* and a significance *S*.

Bayesian Belief Networks: Bayesian belief networks (BBN) are graphical representations of probability distributions, derived from co-occurrence counts in the set of data items. Specifically, a BBN is a directed, acyclic graph, where the nodes represent attribute variables and the edges represent probabilistic dependencies between the attribute variables. Associated with each node are conditional probability distributions that describe the relationships between the node and its parents.

Genetic algorithms / Evolutionary Programming: Genetic algorithms and evolutionary programming are algorithmic optimization strategies that are inspired by the principles observed in natural evolution. Of a collection of potential problem solutions that compete with each other, the best solutions are selected and combined with each other. In doing so, one expects that the overall goodness of the solution set will become better and better, similar to the process of evolution of a population of organisms. Genetic algorithms and evolutionary programming are used in data mining to formulate hypotheses about dependencies between variables, in the form of association rules or some other internal formalism.

Fuzzy Sets: Fuzzy sets form a key methodology for representing and processing uncertainty. Uncertainty arises in many forms in today's databases: imprecision, non-specificity, inconsistency, vagueness, etc. Fuzzy sets exploit uncertainty in an attempt to make system complexity manageable. As such, fuzzy sets constitute a powerful approach to deal not only with incomplete, noisy or imprecise data, but may also be helpful in developing uncertain models of the data that provide smarter and smoother performance than traditional systems. Since fuzzy systems can tolerate uncertainty and can even utilize language-like vagueness to smooth data lags, they may offer robust, noise tolerant models or predictions in situations where precise input is unavailable or too expensive.

Rough Sets: A rough set is defined by a lower and upper bound of a set. Every member of the lower bound is a certain member of the set. Every non-member of the upper bound is a certain non-member of the set. The upper bound of a rough set is the union between the lower bound and the so-called boundary region. A member of the boundary region is possibly (but not certainly) a member of the set. Therefore, rough sets may be viewed as fuzzy sets with a three-valued membership function (yes, no, perhaps). Like fuzzy sets, rough sets are a mathematical concept dealing with uncertainty in data ([42]). Also like fuzzy sets, rough sets are seldom used as a stand-alone solution; they are usually combined with other methods such as rule induction, classification, or clustering methods.

A discussion of the advantages and disadvantages for the different approaches with respect to data mining as well as pointers to introductory readings are given in [24].

3. INVESTIGATION OF KNOWLEDGE DISCOVERY AND DATA MINING TOOLS USING A FEATURE CLASSIFICATION SCHEME

In this section we first provide a feature classification scheme to study knowledge discovery and data mining tools. We then apply this scheme to review existing tools that are currently available, either as a research prototype or as a commercial product. Although not exhaustive, we believe that the reviewed products are representative for the current status of technology.

As discussed in Section 2, knowledge discovery and data mining tools require a tight integration with database systems or data warehouses for data selection, preprocessing, integrating, transformation, etc. Not all tools have the same database characteristics in terms of data model, database size, queries supported, etc. Different tools may perform different data mining tasks and employ different methods to achieve their goals. Some may require or support more interaction with the user than the

other. Some may work on a stand-alone architecture while the other may work on a client/server architecture. To capture all these differences, we propose a feature classification scheme that can be used to study knowledge discovery and data mining tools. In this scheme, the tools' features are classified into three groups called general characteristics, database connectivity, and data mining characteristics which are described below.

3.1.1 General Characteristics (Table 3.1)

Product: Name and vendor of the software product.

Production Status: Status of product development. P=Commercial strength product, A=Alpha, B=Beta, R=Research Prototype.

Legal Status: PD=Public Domain, F=Freeware, S=Shareware, C=Commercial.

Acad. Licensing: Specifies for commercial products, if there is a special free or reduced-cost academic licensing available.

Demo: Specifies if there is a demo version available. D=Demo version available for download on the internet, R=Demo available on request, U=Unknown.

Architecture: The computer architecture on which the software runs. S=Standalone, C/S=Client/Server, P=Parallel Processing.

Operating Systems: Lists the operating systems for which run time version of the software can be obtained.

3.1.2 Database Connectivity (Table 3.2)

Data sources: Specifies possible formats for the data that is to be analyzed. T=Ascii text files, D=Dbase files, P=Paradox files, F=Foxpro files, Ix=Informix, O=Oracle, Sy=Sybase, Ig=Ingres, A=MS Access, OC=Open database connection (ODBC), SS=MS SQL Server, Ex=MS Excel, L=Lotus 1-2-3.

DB Conn.: Type of database connection. Onl.=Online: Queries are run directly against the database and may run concurrently with other transactions. Offl.=Offline: The analysis is performed with a snapshot of the data source.

Size: The maximum number of records the software can comfortably handle. S=Small (up to 10.000 records), M=Medium (10.000 to 1.000.000 records), L=Large (more than 1.000.000 records).

Model: The data model for the data to be analyzed. R=Relational, O = Object Oriented, 1=One table.

Attributes: The type of the attributes the software can handle. Co=Continuous, Ca=Categorical (discrete numerical values), S=Symbolic.

Queries: Specifies how the user can formulate queries against the knowledge base and direct the discovery process. S=Structured query language (SQL or derivative), Sp.=an application specific interface language, G=Graphical user interface, N=Not applicable, U=Unknown.

3.1.3 Data Mining Characteristics (Table 3.3)

Discovery Tasks: Knowledge discovery tasks that the product is intended for. Pre.=Data Preprocessing (Sampling, Filtering), P=Prediction, Regr=Regression, Cla=Classification, Clu=Clustering, A=Link Analysis (Associations), Vis=Model Visualization, EDA = Exploratory Data Analysis.

Discovery Methodology: Type of methods used to discover the knowledge. NN=Neural Networks, GA=Genetic Algorithms, FS=Fuzzy Sets, RS=Rough Sets, St.=Statistical Methods, DT=Decision Trees, RI=Rule Induction, BN=Bayesian Networks, CBR = Case Based Reasoning.

Human Interaction: Specifies how much human interaction with the discovery process is required and/or supported. A=Autonomous, G=Human guided discovery process, H=Highly Interactive.

In tables 3.1, 3.2, and 3.3, we apply our proposed feature classification scheme to study 43 existing knowledge discovery and data mining tools. Although the importance of the ability to access a wide variety of data sources is widely recognized by now, table 3.2 shows that the majority of currently available tools still support only a small number of data formats. Surprisingly few tools offer the ability to analyze several tables simultaneously. However, almost all of the reviewed products can analyze continuous as well as discrete and symbolic attribute types. Mostly this is implemented by conversion of attribute types; i.e. transforming symbolic variables into numerical ones or splitting of continuous attribute ranges into intervals.

From table 3.3 we can observe that most of the tools employ "standard" data mining techniques like rule induction, decision trees, and statistical methods. Techniques from other promising fields like fuzzy and rough sets or genetic algorithms have only begun yet to find their way into knowledge discovery software.

Detailed descriptions for each software product are provided in [24].

4. CONCLUSIONS AND FUTURE RESEARCH

Knowledge discovery can be broadly defined as the automated discovery of novel and useful information from commercial databases. Data mining is one step at the core of the knowledge discovery process, dealing with the extraction of patterns and relationships from large amounts of data.

Today, most enterprises are actively collecting and storing large databases. Many of them have recognized the potential value of these data as an information source for making business decisions. The dramatically increasing demand for better decision support is answered by an extending availability of knowledge discovery and data mining products, in the form of research prototypes developed at various universities as well as software products from commercial vendors. In this paper, we provide an overview of common knowledge discovery tasks, approaches to solve these tasks, and available software tools employing these approaches.

However, despite its rapid growth, KDD is still an emerging field. The development of successful data mining applications still remains a tedious process ([21]). The following is a (naturally incomplete) list of issues that are unexplored or at least not satisfactorily solved yet:

Integration of different techniques. Currently available tools deploy either a single technique or a limited set of techniques to carry out data analysis. From section 2 it follows immediately that there is no best technique for data analysis. The issue is therefore not which technique is better than another, but rather which technique is suitable for the problem at hand. A truly useful tool

(continued after results tables…)

Product	Prod. Status	Legal Status	Acad. Lic.	Demo	Arch.	Operating System							
						MF	Unix	Mac	Dos	W3.x	W9x	WNT	OS/2
Alice (Isoft) [29]	P	C	N	D	S						x	x	
AutoClass C (Nasa) [55]	P	PD	-	D	S, P		x		x				
Bayesian Knowl. Disc. (Open U.) [48]	B	F	-	D	S		x	x					
BrainMaker (Cal. Sc. Software) [22]	P	C	N	N	S			x	x	x	x	x	
Brute (Univ. of Washington) [52]	P	C	Y	N	S		x						
BusinessMiner (Business Objects) [11]	P	C	N	N	S					x	x	x	
Claudien (K. U. Leuven) [15]	P	C	Y	D	S		x						
Clementine (Integral Solutions Ltd.) [28]	P	C	N	N	S		x					x	
CN2 (Univ. of Texas) [12]	R	PD	-	D	S		x						
CrossGraphs (Belmont Research) [9]	P	C	N	N	S		x	x		x	x	x	
Cubist (RuleQuest) [51]	P	C	D	D	S		x				x	x	
C5.0 (RuleQuest) [47]	P	C	D	D	S		x			x	x	x	
Darwin (Thinking Machines) [7]	P	C	N	N	C/S, P, S		x						
Data Surveyor (Data Destilleries) [14]	P	C	S	R	C/S, P		x	x		x	x	x	x
DBMiner (SFU) [30]	R	C	S	N	C/S		x				x	x	
Delta Miner (Bissantz) [5]	P	C	S	D	C/S, S						x	x	
Decision Series (NeoVista) [39]	P	C	N	N	C/S, P		x						x
DI Diver (Dimensional Insight) [17]	P	C	N	D	C/S, S		x	x		x	x	x	
Ecobweb (Tel Aviv Univ.) [50]	P	PD	-	D	S		x						
IDIS (Information Discovery) [27]	P	C	N	N	C/S, S, P	x	x	x		x	x	x	x
IND (Nasa) [38]	P	C	D	N	S		x						
Intelligent Miner (IBM) [25]	P	C	S	R	C/S, S, P	x	x			x	x	x	x
KATE-Tools (AcknoSoft) [2]	P	C	N	N	C/S, S		x	x		x	x	x	
Kepler (GMD) [59]	R	C	S	R	S		x				x		
KnowledgeSEEKER (Angoss) [3]	P	C	N	D	C/S, S		x			x	x	x	
MineSet (Silicon Graphics) [10]	P	C	S	N	C/S, S		x						
MLC++ (Silicon Graphics) [31]	P	F	-	D	S		x					x	
Mobal (GMD) [54]	P	F	-	D	S		x						

Table 3.1: General Product Characteristics - Part 1

Product	Prod. Status	Legal Status	Acad. Lic.	Demo	Arch.	Operating System							
						MF	Unix	Mac	Dos	W3.x	W9x	WNT	OS/2
ModelQuest (AbTech) [1]	P	C	N	N	S		x				x	x	
MSBN (Microsoft) [36]	P	F	-	D	S						x	x	
MVSP (KCS) [32]	P	S	D	D	S				x	x	x	x	
PolyAnalyst (Megaputer) [34]	P	C	F	D	C/S, S						x	x	x
PVE (IBM) [26]	P	C	N	N	S		x						
Q-Yield (Quadrillion) [46]	P	C	F	D	S						x	x	
Ripper (AT&T) [13]	P	C	F	D	S		x						
Rosetta (NTNU) [40]	R	C	S	D	S						x	x	
Rough Enough (Troll Data) [6]	R	F	-	D	S				x	x	x	x	
Scenario (Cognos) [23]	P	C	N	R	S						x	x	
Sipina-W (Univ. of Lyon) [57]	p	S	-	D	S					x			
SuperQuery (AZMY) [4]	P	C	N	D	S						x	x	
ToolDiag (UFES) [49]	P	F	-	D	S		x		x				
Weka (Univ. of Waikato) [58]	P	C	F	D	S		x						
WizRule (WizSoft) [37]	P	C	N	D	S						x	x	

Table 3.1: General Product Characteristics - Part 2

Product: Name and vendor of the software product.

Production Status: Status of product development. P = Commercial strength product, A = Alpha, B= Beta, R = Research Prototype.

Legal Status: PD = Public Domain, F = Freeware, S = Shareware, C = Commercial.

Acad. Licensing: Specifies for commercial products, if there is a special free or reduced-cost academic licensing available.

Demo: Specifies if there is a demo version available. D = Demo version available for download on the internet, R = Demo available on request, U = Unknown.

Architecture: The computer architecture on which the software runs. S = Standalone, C/S = Client/Server, P = Parallel Processing.

Operating Systems: Lists the operating systems for which run time version of the software can be obtained.

Product	Data sources													DB Conn.		Size	Mod.	Attributes			Queries
	T	D	P	F	Ix	O	Sy	Ig	A	OC	SS	Ex	L	Onl.	Offl.			Co	Ca	S	
Alice (Isoft) [29]	x	x	x	x		x			x	x		x	x		x	L	1	x	x	x	S, G
AutoClass C (Nasa) [55]	x	x													x	L	1	x	x	x	N
Bayesian Knowl. Disc. (Open U.) [48]	x														x	L	1		x		N
BrainMaker (Cal. Sc. Software) [22]	x	x										x	x		x	S	1	x	x		N
Brute (Univ. of Washington) [52]	x														x	S	1	x	x	x	N
BusinessMiner (Business Objects) [11]	x											x			x	M	R	x	x	x	S, G
Claudien (K. U. Leuven) [15]	x														x	S	R		x	x	U
Clementine (Integral Solutions Ltd.) [28]	x				x	x	x	x		x		x	x	x	x	M	R	x	x	x	G
CN2 (Univ. of Texas) [12]	x														x	S	1		x	x	N
CrossGraphs (Belmont Research) [9]	x	x		x		x				x				x	x	M	R	x	x	x	Sp.
Cubist (RuleQuest) [51]	x														x	M	1	x	x	x	N
C5.0 (RuleQuest) [47]	x														x	M	1	x	x	x	N
Darwin (Thinking Machines) [7]	x	x	x	x		x	x							x		L	R	x	x	x	S
Data Surveyor (Data Destilleries) [14]		x	x	x		x	x			x				x		L	R	x	x	x	S
DBMiner (SFU) [30]										x				x		L	R	x	x	x	S, G
Delta Miner (Bissantz) [5]										x		x			x	L	R	x	x	x	S, G
Decision Series (NeoVista) [39]	x	x			x	x	x			x				x		L	R	x	x	x	S
DI Diver (Dimensional Insight) [17]	x									x				x	x	M	R	x	x	x	S, G
Ecobweb (Tel Aviv Univ.) [50]	x														x	S	1	x	x	x	N
IDIS (Information Discovery) [27]	x	x				x	x				x			x	x	L	R	x	x	x	S, G
IND (Nasa) [38]	x														x	S	1	x	x	x	N
Intelligent Miner (IBM) [25]	x	x				x	x							x	x	L	R	x	x	x	S,G
KATE-Tools (AcknoSoft) [2]	x					x			x	x	x			x	x	M	R	x	x	x	S, G
Kepler (GMD) [59]	x														x	M	R	x	x	x	S, G
KnowledgeSEEKER (Angoss) [3]	x	x	x						x	x	x	x	x		x	L	1	x	x	x	S, G
MineSet (Silicon Graphics) [10]	x				x	x	x							x	x	L	R	x	x	x	S, G
MLC++ (Silicon Graphics) [31]	x														x	L	R	x	x	x	N
Mobal (GMD) [54]	x														x	S	R		x	x	S, G

Table 3.2: Database Connectivity - Part 1

Product	Data sources													DB Conn.		Size	Mod.	Attributes			Queries
	T	D	P	F	Ix	O	Sy	Ig	A	OC	SS	Ex	L	Onl.	Offl.			Co	Ca	S	
ModelQuest (AbTech) [1]	x														x	L	R	x	x	x	G
MSBN (Microsoft) [36]	x														x	S	R	x	x	x	G
MVSP (KCS) [32]	x	x	x									x	x		x	M	1	x	x		N
PolyAnalyst (Megaputer) [34]	x	x	x	x	x	x	x	x		x	x	x			x	M	R	x	x	x	G
PVE (IBM) [26]	x	x													x	M	1	x	x	x	U
Q-Yield (Quadrillion) [46]	x									x				x		S	R	x	x	x	S
Ripper (AT&T) [13]	x														x	L	1	x	x	x	N
Rosetta (NTNU) [40]	x									x					x	S	R	x	x	x	S, G
Rough Enough (Troll Data) [6]	x		x											x	x	S	R	x	x	x	S
Scenario (Cognos) [23]	x	x	x	x								x		x	x	M	R	x	x		S, G
Sipina-W (Univ. of Lyon) [57]	x	x	x										x		x	M	1	x	x	x	G
SuperQuery (AZMY) [4]	x	x	x	x					x	x		x	x	x		M	R	x	x	x	G, A
ToolDiag (UFES) [49]	x														x	S	1	x			N
Weka (Univ. of Waikato) [58]	x														x	M	R	x	x	x	N
WizRule (WizSoft) [37]	x	x	x							x				x		M	1	x	x	x	N

Table 3.2: Database Connectivity - Part 2

Data sources: Specifies possible formats for the data which is to be analyzed. T = Ascii text files, D = Dbase files, P = Paradox files, F = Foxpro files, Ix = Informix, O = Oracle, SY = Sybase, Ig = Ingres, A = MS Access, OC = Open database connection (ODBC), SS = MS SQL Server, Ex = MS Excel, L = Lotus 123.

DB Conn.: Type of database connection. Onl. = Online: Queries are run directly against the database and may run concurrently with other transactions. Offl. = Offline: The analysis is performed with a snapshot of the data source.

Size: The maximum number of records the software can comfortably handle, S = Small (up to 10.000 records), M = Medium (10.000 to 1.000.000 records), L = Large (more than 1.000.000 records).

Model: The data model for the data to be analyzed. R = Relational, O = Object Oriented, N = Not applicable (the software takes just one table).

Attributes: The type of the attributes the software can handle. Co = Continuous, Ca = Categorical (discrete numerical values), S = Symbolic.

Queries: Specifies how the user can formulate queries against the knowledge base and direct the discovery process. S = Structured query language (SQL or derivative), Sp. = an application specific interface language, G = Graphical user interface, N = Not applicable, U = Unknown.

Product	Tasks								Methods									Interaction		
	Pre.	P	Regr	Cla	Clu	A	Vis	EDA	NN	GA	FS	RS	St.	DT	RI	BN	CBR	A	G	H
Alice (Isoft) [29]				x			x							x				x	x	
AutoClass C (Nasa) [55]	x			x	x								x					x		
Bayesian Knowl. Disc. (Open U.) [48]				x		x										x		x		
BrainMaker (Cal. Sc. Software) [22]		x	x	x					x									x	x	
Brute (Univ. of Washington) [52]				x		x									x			x		
BusinessMiner (Business Objects) [11]		x		x		x	x							x						x
Claudien (K. U. Leuven) [15]				x		x									x			x		
Clementine (Integral Solutions Ltd.) [28]	x	x	x	x	x	x	x	x	x					x	x					x
CN2 (Univ. of Texas) [12]				x		x									x			x		
CrossGraphs (Belmont Research) [9]	x						x	x					x							x
Cubist (RuleQuest) [51]		x	x		x										x			x		
C5.0 (RuleQuest) [47]				x										x	x			x		
Darwin (Thinking Machines) [7]	x	x	x	x	x	x	x	x	x	x				x				x	x	x
Data Surveyor (Data Destilleries) [14]	x	x	x	x	x	x	x	x	x				x	x	x	x				x
DBMiner (SFU) [30]	x	x	x	x		x	x						x		x					x
Delta Miner (Bissantz) [5]	x	x	x	x		x	x	x	x				x						x	
Decision Series (NeoVista) [39]	x	x	x	x	x	x			x				x	x	x				x	
DI Diver (Dimensional Insight) [17]	x	x		x			x	x					x							x
Ecobweb (Tel Aviv Univ.) [50]		x		x	x								x						x	
IDIS (Information Discovery) [27]		x	x	x	x	x	x				x				x			x	x	
IND (Nasa) [38]		x	x	x										x				x		
Intelligent Miner (IBM) [25]	x	x	x	x					x					x						x
KATE-Tools (AcknoSoft) [2]	x	x	x	x			x							x			x			x
Kepler (GMD) [59]	x	x	x	x	x	x	x	x	x				x	x	x		x	x	x	x
KnowledgeSEEKER (Angoss) [3]		x		x			x							x	x				x	
MineSet (Silicon Graphics) [10]	x	x		x	x	x	x	x					x	x	x					x
MLC++ (Silicon Graphics) [31]	x	x	x	x	x	x	x						x	x	x		x		x	
Mobal (GMD) [54]		x				x	x								x					x

Table 3.3: Data Mining Characteristics - Part 1

Product	Tasks								Methods									Interaction		
	Pre.	P	Regr	Cla	Clu	A	Vis	EDA	NN	GA	FS	RS	St.	DT	RI	BN	CBR	A	G	H
ModelQuest (AbTech) [1]	x	x	x	x	x		x		x				x	x			x	x	x	
MSBN (Microsoft) [36]		x				x	x									x				x
MVSP (KCS) [32]		x		x	x								x							x
PolyAnalyst (Megaputer) [34]	x	x	x	x	x	x	x			x			x		x			x	x	
PVE (IBM) [26]	x	x	x				x	x					x							x
Q-Yield (Quadrillion) [46]						x	x						x		x			x	x	
Ripper (AT&T) [13]		x	x			x									x			x		
Rosetta (NTNU) [40]	x	x		x		x				x		x	x		x				x	
Rough Enough (Troll Data) [6]	x					x						x			x					x
Scenario (Cognos) [23]	x	x				x							x	x						x
Sipina-W (Univ. of Lyon) [57]	x	x	x	x		x	x							x	x			x	x	
SuperQuery (AZMY) [4]	x	x		x		x	x	x					x		x			x		x
ToolDiag (UFES) [49]		x		x					x				x				x			x
Weka (Univ. of Waikato) [58]	x	x	x	x	x	x							x	x	x		x		x	
WizRule (WizSoft) [37]						x									x			x		

Table 3.3: Data Mining Characteristics - Part 2

Discovery Tasks: Knowledge discovery tasks that the product is intended for. Pre. = Data Preprocessing (Sampling, Filtering), P = Prediction, Regr. = Regression, Cla = Classification, Clu = Clustering, A = Link Analysis (Associations), Vis = Model Visualization, EDA = Exploratory Data Analysis.

Discovery Methodology: Types of methods used to discover the knowledge. NN = Neural Networks, GA = Genetic Algorithms, FS = Fuzzy Sets, RS = Rough Sets, St. = Statistical Methods, DT = Decision Trees, RI = Rule Induction, BN = Bayesian Networks, CBR = Case Based Reasoning.

Human Interaction: Specifies how much human interaction with the discovery process is required and/or supported. A = Autonomous, G = Human guided discovery process, H = Highly Interactive.

has to provide a wide range of different techniques for the solution of different problems.

Extensibility. This is another consequence from the fact that different techniques outperform each other for different problems. With the increasing number of proposed techniques as well as reported applications, it becomes clearer and clearer that any fixed arsenal of algorithms will never be able to cover all arising problems and tasks. It is therefore important to provide an architecture that allows for easy synthesis of new methods, and adaptation of existing methods with as little effort as possible.

Seamless integration with databases. Currently available data analysis products generally fall into two categories. The first is drill-down analysis and reporting, provided by vendors of RDBMS's, sometimes in association with on-line analytical processing (OLAP) vendors. These systems provide a tight connection with the underlying database and usually deploy the processing power and scalability of the DBMS. They are also restricted to testing user provided hypotheses, rather than automatically extracting patterns and models. The second category consists of (usually stand-alone) pattern discovery tools, which are able to autonomously detect patterns in the data. These tools tend to access the database offline; that is, data is extracted from the database and fed into the discovery engine. Many tools even rely on keeping all their data in main memory, thus lacking scalability and, therefore, the ability to handle real world problems. Additionally, these tools are often insufficiently equipped with data processing capabilities, leaving the data preprocessing solely to the user. This can result in time-consuming repeated export-import processes.

Support for both analysis experts and novice users. There are usually three stages at deploying KDD technology in an organization:

1) The potential of KDD is discovered. First naive studies are performed, often by external consultants (which are data mining specialists).
2) Once the profitability of KDD is proven, it is used on a regular basis to solve business problems. Users usually are teams of analysis experts (with expertise in KDD technology) and domain experts (with extensive knowledge of the application domain).
3) Fully exploitation of KDD technology within the organization. End users are enabled to perform their own analysis according to their individual needs. Although widely still a vision, the necessity for this stage is clearly recognized.

Obviously the different users at these stages have different demands and also bring different prerequisites. Most of the available tools are aimed at analysis experts, requiring an unaffordable amount of training before being useful to novice end users. Typical end users are for example marketers, engineers or managers. These users are less skilled in complex data analysis and have less knowledge of the nature of the data available, but have a thorough understanding of their occupation domain. Furthermore, they are usually not interested in using advanced powerful technology, but only in getting clear, rapid answers to their everyday business questions.

End users need simple-to-use tools that efficiently solve their business problems. Existing software packages lack sufficient support for both directing the analysis process and presenting the analysis results in a user-understandable manner. If not, they are restricted to a very limited set of techniques and problems. Optimally, a better usability by novice users would have to be achieved without giving up other desirable features such as flexibility and/or analysis power.

Managing changing data. In many applications, including the vast variety of nearly all business problems, the data is not stationary, but rather changing and evolving. This changing data may make previously discovered patterns invalid and hold new ones instead. Currently, the only solution to this problem is to repeat the same analysis process (which is also work-intensive) in periodic time intervals. There is clearly a need for incremental methods that are able to update changing models, and for strategies to identify and manage patterns of temporal change in knowledge bases.

Non-standard data types. Today's databases do not contain only standard data such as numbers and strings but also large amounts of nonstandard and multimedia data, such as free-form text, audio, image and video data, temporal, spatial and other data types. Those data types contain special patterns, which can not be handled well by the standard analysis methods. Therefore, these applications require special, often domain-specific, methods and algorithms.

While there are fundamental problems that remain to be solved, there have also been numerous significant success stories reported, and the results and benefits are impressive ([53]). Although the current methods still rely on fairly simple approaches with limited capabilities, reassuring results have been achieved, and the benefits of KDD technology have been convincingly demonstrated in the broad range of application domains. The combination of urgent practical needs and the strong research interests lets us also expect a future healthy grow of the field, drawing KDD tools into the mainstream of business applications.

5. References

[1] AbTech Corp. ModelQuest 2.0. http://www.abtech.com. Charlottesville, VA, 1997.

[2] Acknosoft Inc. KATE-Tools 6.0. http://www.acknosoft.com. Palo Alto, CA, 1999.

[3] Angoss Software Ltd. KnowledgeSEEKER 4.4. http://www.angoss.com. Guildford Surrey, UK, 1998.

[4] AZMY Thinkware Inc. SuperQuery 1.50. http://www.azmy.com. Fort Lee, NJ, 1997.

[5] Bissantz Küppers & Company GmbH. Delta Miner 3.5. http://www.bissantz.de, Erlangen, Germany, 1998.

[6] Bjorvand, A.T. Rough Enough -- Software Demonstration. 15th IMACS World Congress on Scientific Computation, Modelling and Applied Mathematics. Berlin, Germany, August 24-29, 1997.

[7] Bourgoin, M.O., and Smith, S.J. Big Data -- Better Returns, Leveraging Your Hidden Data Assets to Improve ROI. In Freedman et al (eds.), Artificial Intelligence in the Capital Markets, Probus Publishing Company, 1995.

[8] Brachman, R., and Anand, T. The process of knowledge discovery in databases: A human-centered Approach. In

Fayyad, U., Piatetsky-Shapiro, G., Amith, Smyth, P., and Uthurusamy, R. (eds.), Advances in Knowledge Discovery and Data Mining, MIT Press, Cambridge, 1996.

[9] Brooks, P. Visualizing Data -- Sophisticated Graphic Visualization and Development Tools Tailored for Business Applications. In DBMS, Miller Freeman Inc., San Mateo, CA, August 1997.

[10] Brunk, C., Kelly, J., and Kohavi, R. MineSet: An Integrated System for Data Mining. In Proceedings of the The Third International Conference on Knowledge Discovery and Data Mining, August 1997. (Can be retrieved from http://robotics.stanford.edu/users/ronnyk/ronnyk-bib.html).

[11] Business Objects S.A. BusinessMiner 4.1. http://www.businessobjects.com, San Jose, CA, 1998.

[12] Clark, P., and Boswell, R. Rule induction with CN2: Some recent improvements. In Kodratoff, Y. (ed.), Machine Learning - EWSL-91, Springer-Verlag, Berlin, 151-163, 1991.

[13] Cohen, W. Fast effective rule induction. Proceedings of the Twelfth International Conference on Machine Learning, Lake Tahoe, California, 1995.

[14] Data Destilleries Inc. Online Data Mining (tm) -- Data Destilleries' vision on data mining technology. White Paper DD-R9704, Amsterdam, May 1997.

[15] Dehaspe, L., Van Laer, W., and De Raedt, L. Claudien, the Clausal Discovery Engine -- User's Guide 3.0. Technical Report CW 239, Department of Computing Science, K.U.Leuven, 1996.

[16] Demuth, H., and Beale, M. The Neural Network Toolbox for MATLAB. Mathworks Inc., Nattick, MA, 1996.

[17] Dimensional Insight Inc. DI Diver 3.0. http://www.dimins.com, 1999.

[18] Exclusive Ore Inc. The Exclusive Ore Internet Site, http://www.xore.com, 1999.

[19] Fayyad, U. Data Mining and Knowledge Discovery: Making Sense Out of Data. IEEE Expert, v. 11, no. 5, pp. 20-25, October 1996.

[20] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. From Data Mining to Knowledge Discovery: An Overview. In Fayyad, U., Piatetsky-Shapiro, G., Amith, Smyth, P., and Uthurusamy, R. (eds.), Advances in Knowledge Discovery and Data Mining, MIT Press, 1-36, Cambridge, 1996.

[21] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communications of the ACM, v. 39, no. 11, pp. 27-34, November 1996.

[22] Flori, R.D. Product Review: BrainMaker Professional Neural Network Simulation. Computerized Investing, American Association Of Individual Investors, v. 11, no. 1, January/February 1992.

[23] Gilliland, S. Scenario 1.0: Mine Your Data for Statistical Gems. Computer Shopper, Ziff-Davis Publishing Company, October 1997.

[24] Goebel, M., and Gruenwald, L. A Survey of Knowledge Discovery and Data Mining Tools. Technical Report, University of Oklahoma, School of Computer Science, Norman, OK, February 1998.

[25] IBM Corporation. Intelligent Miner 2.1. http://www.software.ibm.com/data/intelli-mine, 1998.

[26] IBM Corporation. PVE 1.0. http://www.ibm.com/news/950203/pve-01.html, 1995.

[27] Information Discovery Inc. The IDIS Information Discovery System. http://www.datamining.com, 1997.

[28] Integral Solutions Ltd. Clementine 5.0. http://www.isl.co.uk/clem.html, 1998

[29] Isoft SA. Alice 5.1. http://www.alice-soft.com, 1998.

[30] Kamber, M., Han, J., and Chiang, J.Y. Using Data Cubes for Metarule-Guided Mining of Multi-Dimensional Association Rules. Technical Report U-SFraser-CMPT-TR: 1997-10, Simon Fraser University, Burnaby, May 1997.

[31] Kohavi, R., Sommerfield, D., and Dougherty, J. Data Mining using MLC++: A Machine Learning Library in C++. Tools with AI '96, 234-245, November 1996.

[32] Kovach Computing Services. MVSP 3.0. http://www.kovcomp.co.uk/mvsp.html, 1997.

[33] Lippmann, R.P. An Introduction to Computing with Neural Nets. IEEE ASSP Magazine, 4-22, April 1987.

[34] Megaputer Intelligence Ltd. PolyAnalyst 3.5 http://www.megaputer.com, 1998.

[35] Meta Group Inc. Data Mining: Trends, Technology, and Implementation Imperatives. Stamford, CT, Februay 1997.

[36] Microsoft Corp. MSBN 1.0, http://www.research.microsoft.com/research/dtg/msbn, 1996.

[37] Nicolaisen, N. WizRule may be the key to avoiding database disasters. Computer Shopper, Ziff-Davis Publishing Company, L.P., 588-591, November 1995.

[38] NASA COSMIC. IND 1.0. University of Georgia, NASA COSMIC Department, http://www.cosmic.uga.edu, 1992.

[39] NeoVista Solutions Inc. Decision Series 3.0, http://www.neovista.com, 1998.

[40] Øhrn, A., Komorowski, J., Skowron, A., and Synak, P. The Design and Implementation of a Knowledge Discovery Toolkit Based on Rough Sets -- The ROSETTA System. In Polkowski, L., and Skowron, A. (eds.), Rough Sets in Knowledge Discovery, Physica Verlag, 1998.

[41] Oracle Inc. Discoverer 3.0 User's Guide, 1997.

[42] Pawlak, Z. Rough Sets: Theoretical Aspects of Reasoning About Data. Kluwer, Boston, 1991.

[43] Price Waterhouse LLP. Geneva V/T 3.2. Geneva (Switzerland), December 1997.

[44] Piatetsky-Shapiro, G. The Knowledge Discovery Mine, http://www.kdnuggets.com, 1999.

[45] Pryke, A. The Data Mine. http://www.cs.bham.ac.uk/~anp/, 1999.

[46] Quadrillion Corp. Q-Yield 4.0. http://www.quadrillion.com, 1999.

[47] Quinlan, J.R., C4.5 -- Programs for Machine Learning. Morgan Kaufmann Publishers, San Francisco, CA, 1993.

[48] Ramoni, M., and Sebastiani, P. Learning Bayesian Networks from Incomplete Databases. In Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufman, San Mateo, CA, 1997.

[49] Rauber, T.W. ToolDiag 2.1. Universidade Federal do Espirito Santo, Departamento de Informatica, http://www.inf.ufes.br/~thomas/www/home/tooldiag.html, 1998.

[50] Reich, Y. Building and Improving Design Systems: A Machine Learning Approach. PhD thesis, Department of Civil Engineering, Carnegie Mellon University, Pittsburgh, PA, 1991. (Available as Technical Report EDRC 02-16-91 from the Engineering Design Research Center at CMU).

[51] RuleQuest Research Inc. Cubist 1.06a. http://www.rulequest.com, 1999.

[52] Segal, R., and Etzioni, O. Learning decision lists using homogeneous rules. In Proceedings of the Twelfth National Conference on Artificial Intelligence, July 1994.

[53] Simoudis, E. Reality Check for Data Mining, IEEE Expert, v. 11, no. 5, pp. 20-25, October 1996.

[54] Sommer, E., Emde, W., Kietz, J., Morik, K., and Wrobel, S. MOBAL 2.2 User Guide. Arbeitspapiere der GMD no. 777, St. Augustin (Germany), 1993.

[55] Stutz, J., and Cheeseman, P. AutoClass -- a Bayesian Approach to Classification. In Skilling J., and Sibisi, S. (eds.), Maximum Entropy and Bayesian Methods, Kluwer Academic Publishers, 1995.

[56] Thearling, K. Data Mining and Database Marketing WWW Pages. http://www.santafe.edu/~kurt/dmvendors.shtml, 1998.

[57] University of Lyon. Sipina-W 2.5. Laboratoire E.R.I.C., http://eric.univ-lyon2.fr/~ricco/sipina.html, 1996.

[58] Waikato ML Group. User Manual Weka: The Waikato Environment for Knowledge Analysis. Department of Computer Science, University of Waikato (New Zeaand), June 1997.

[59] Wrobel, S., Wettschereck, D., Sommer, E., and Emde, W. Extensibility in data mining systems. In Simoudis, E., Han, J.W., and Fayyad, U. (eds.), Proc. 2nd International Conference On Knowledge Discovery and Data Mining, 214-219, AAAI Press, Menlo Park, CA, August 1996.

About the authors:

Michael Goebel received his BS degree in Computer Science from the University of Braunschweig, Germany and MS degree in Computer Science from the University of Oklahoma in 1998. He is currently a doctoral candidate in the PhD program in Computer Science at the University of Auckland, New Zealand. His primary research interests are in Machine Learning, Data Mining, and Reasoning under Uncertainty.

Dr. Le Gruenwald is a Presidential Professor and an Associate Professor in the School of Computer Science at The University of Oklahoma. She received her Ph.D. in Computer Science from Southern Methodist University in 1990, MS in Computer Science from the University of Houston in 1983, and BS in Physics from the University of Saigon, Vietnam in 1978. She was a Software Engineer at WRT, a Lecturer in the Computer Science and Engineering Department at Southern Methodist University, and a Member of Technical Staff in the Database Management Group at the Advanced Switching Laboratory of NEC America. Her major research interests include Real-Time Databases, Object-Oriented Databases, Data Warehouse and Data Mining, Multimedia Databases, and Distributed Mobile Databases. She is a member of ACM, SIGMOD, and IEEE Computer Society.